THIS WEEK

EDITORIALS

BALLOONS The rise and rise of research using our inflatable friends **p.578**

WORLD VIEW Scottish science should take the high road **p.579**

MILKY WAY Dung beetles prefer a starry, starry night **p.580**

Twice the price

Governments and funding agencies must do more to prevent the awarding of grants to research projects with significant overlap.

There is nothing more central to the modern world of international science than the research grant. And with government budgets squeezed, there is nothing more important than making sure that what money remains for project-based science is spent wisely. So scientists everywhere should be disturbed that two separate pieces in *Nature* this week report on the lack of oversight of potential waste and overlap between research grants.

Similarities between a number of US grants were first flagged up by a Comment on bioinformatics research (see page 599). Two reporters then requested more details under the US Freedom of Information Act (see page 588). Just as important as what we found is what we couldn't find.

It turns out that although some individual agencies maintain databases, in most countries — perhaps even all — there is no centralized government-maintained online database of all state-funded research projects. This week's findings come from three US government agencies that do keep such records: the US Army's Congressionally Directed Medical Research Programs, the National Institutes of Health and the National Science Foundation (NSF). There is no reason to think that these agencies are not representative. So the findings, limited though they are, warrant careful attention.

A review of 22 pairs of seemingly similar grant files revealed many that appeared to overlap, with specific aims, hypotheses and methods that contained large sections of duplicated text. Where we saw different text, we were careful to analyse whether it had a central role — for example, whether it showed study of an entirely different protein or nanomaterial by an identical method. In many instances, the different text didn't seem to fully distinguish projects from each other. In some cases, researchers and agencies did provide explanations of why seemingly similar grants did not overlap, and these are given in our News story. But the exercise nevertheless exposed some loopholes.

First, checks on overlap are mostly trust-based. The responsibility lies with researchers and institutions to declare when they have been awarded similar grants. Yet some that we reviewed apparently had not done so, or not in a timely fashion. Similarly, researchers sometimes declared "overlap: none" between applications when to us — and sometimes even to agency staff — it seems that there was some overlap. Although much of science is trust-based, there is no reason, with the advent of text-similarity software and electronic databasing, for agencies not to be proactive (in the way the bioinformaticians who prepared the Comment piece were) and ask for more original documentation when large segments of grants seem identical. Indeed some officials, we could see from the files, are already doing this.

Second, concurrent submissions of similar grant applications to US agencies do not have to be declared to every agency involved until funding decisions are made. The NSF does require declaration on submission when applications are identical, but we found that in most cases they were only similar. It is worth considering whether all submissions should be declared up front, in the same way that college and graduate-school applications in the United States and the United Kingdom include information on all applications made by a student. This might help reviewers to better understand each researcher's range of interests, as well as helping agencies to avoid overlap. Agencies should adopt and adapt the NSF checkbox to applications so that instead of asking about duplicate proposals under submission it asks 'do you have any grant applications (submitted or funded) that may overlap with this one?'. If selected, this would trigger a more detailed review.

> ***"There is no reason, with the advent of text-similarity software, for agencies not to be proactive."***

More importantly, agencies worldwide should also follow the example of the three that we examined and create databases of grant funding online, where past and current awards can be easily found by scientific search terms, researchers' names, institution, city and agency. Having created such databases, funding agencies should maintain them.

The US Department of Energy recently took down a useful project database from its website, it says, to save money. But as this information increasingly already exists in-house, the costs of making it public should be modest. The benefit would be that researchers, and others, can see quickly what has been funded and where future efforts are needed. In addition, such a facility would allow the public to understand and scrutinize where its money goes. Of course, the idea of anyone being able to survey funding decisions at a click of a button may make some officials uncomfortable, but those who do a good job to balance and police their portfolios will get the credit they deserve. ■

Change for good

The United States must boost energy spending to make its mark on the climate debate.

Environmentalists lauded US President Barack Obama when he raised the issue of global warming in his second inaugural address on 21 January, but the truth is that he said nothing new. Obama kept it simple, short and vague, discussing climate change as a moral imperative while declaring clean energy a battleground for innovation. It was a generic vision for a pragmatic president, which is to his credit. But if Obama truly wants to leave his mark on the climate debate, he will need to break out of the mould and lay the foundation for something larger.

His initial focus is likely to be a trio of energy decisions, on a pipeline

and a pair of rules for power plants (see page 590). The first decision relates to the Keystone XL pipeline, which would carry oil from the Canadian tar sands to the Gulf Coast refineries. The other two are climate regulations that focus on new and existing power plants. Combined, these two rules could prevent any conventional coal-fired plant from being built in the United States, while giving electricity generation another boost towards using plentiful natural gas.

They give Obama an early opportunity to build some goodwill across the political spectrum. First, the administration should issue strong regulations for power plants and send a message to the coal industry: clean up or fade away. The energy utilities will duly cry foul, but the same companies are already powering down old and inefficient coal-fired power plants in favour of natural-gas plants. Why? Because natural gas is cheap and burns more cleanly than coal, helping companies to meet increasingly stringent air-quality regulations.

Second, regarding the Keystone pipeline, the administration should face down critics of the project, ensure that environmental standards are met and then approve it. As *Nature* has suggested before (see *Nature* **477,** 249; 2011), the pipeline is not going to determine whether the Canadian tar sands are developed or not. Only a broader — and much more important — shift in energy policy will do that. Nor is oil produced from the Canadian tar sands as dirty from a climate perspective as many believe (some of the oil produced in California, without attention from environmentalists, is worse). Tar-sands development raises serious air- and water-quality issues in Canada, but these problems are well outside Obama's jurisdiction.

By approving Keystone, Obama can bolster his credibility within industry and among conservatives. The president can also take advantage of rising domestic oil and gas production to defuse concerns over energy security. And the fact that US emissions are apparently dropping, thanks to the economic crisis and the ongoing shift from coal to gas for electricity generation as well as state and federal policies, further plays into his hands. But all will be for naught unless the president can build on these trends and somehow reset the climate discussion.

The foundation for this re-engagement could be a good old-fashioned strategic research and development (R&D) programme for clean energy. The United States' current US$4-billion energy-research portfolio is not up to the task, and almost everybody recognizes as much. In 2010, the President's Council of Advisors on Science and Technology recommended boosting the federal energy-innovation budget to $16 billion. The Brookings Institution, a Washington DC think tank, has argued that even a small carbon tax could provide up to $30 billion annually for energy research. If these numbers seem high, keep in mind that in fiscal year 2012, the United States spent an estimated $73 billion on defence-related R&D and more than $31 billion on health-related R&D.

"Driving down the cost of low-carbon energy might even unlock political solutions."

These ideas have been floating around in the scientific community for some time. Some extra money will be needed, but organizations such as the Clean Air Task Force, based in Boston, Massachusetts, are looking at ways to better direct energy subsidies and use existing government spending to drive new markets for advanced technologies.

The Obama administration might be able to put the United States on track to meet its Copenhagen commitment to reduce emissions to 17% below 2005 levels by 2020. It can seek immediate climate benefits by pushing international initiatives that reduce emissions of black carbon, methane and other powerful greenhouse gases. But given the current political deadlock over climate regulation on Capitol Hill, Obama must also develop a long game that will help to get the United States, and hopefully the world, to where it wants to be several decades from now. Driving down the cost of low-carbon energy might even unlock political solutions in the future. ■

Inflatable friends

Research balloons have taught us much about the atmosphere, and could now fly into space.

The Swiss physicist Auguste Piccard will be recognizable to anyone who grew up reading the comic-book adventures of Tintin. After spotting Piccard on a Brussels street, the Belgian cartoonist Hergé used his striking appearance as inspiration for Tintin's scientific friend Professor Cuthbert Calculus. But Piccard should also be recognized for his advancement of a scientific platform that remains important today: the research balloon.

Piccard was an inventor and explorer. In 1930, he designed a pressurized steel gondola that could carry passengers and laboratory equipment beneath a balloon. The vehicle would eventually inspire his deep-ocean bathyscaphe, but in 1931 Piccard and his colleague Paul Kipfer used it to explore the atmosphere, reaching 15,785 metres and measuring cosmic rays. It was a fitting experiment: cosmic rays were discovered in 1912 when Austrian physicist Victor Hess carried electrometers to about 5,000 metres in a perilous open basket beneath a balloon.

Balloons have gone higher and farther for science ever since. Just last week, a NASA long-duration balloon broke the record for flight length when it clocked up its 46th day spinning in the high winds and chilly skies above the South Pole. No scientist hangs beneath this one, but the goal remains the same as in Piccard's day. The balloon floats some 39 kilometres up and carries the Super Trans-Iron Galactic Element Recorder, which sifts through high-energy cosmic rays, looking for rare heavy elements.

Balloons could go higher still. This month, NASA raised the prospect that one could be (gently) bolted onto the side of the International Space Station. The agency calls it an expandable activity module; the media used the terms giant space balloon and bouncy castle. Either way, this balloon (expandable activity module) would not simply support science — it could house it. The agency is in talks with the module's developer, Bigelow Aerospace of North Las Vegas, Nevada, as to how it could test the module as a living and working habitat in orbit. If they can repel the radiation and pointy micrometeorites that are a hazard of life in space, then inflatable modules could be used to construct whole space stations. The appeal is obvious: such equipment would be compact and therefore cheap to get off the ground and to construct in orbit.

Balloons have been launched into space before. The twin European Vega missions of the mid-1980s deployed one each to hang in the Venusian sky, where they measured wind speed and cloud density. Balloons have even been used to launch rockets towards space. The 'rockoons' developed by James Van Allen at the University of Iowa in Iowa City in the 1950s were balloons that carried sounding rockets into the atmosphere and then launched them to ever higher altitudes. When the rockets fell back to Earth, they brought hints of layers of trapped radiation beyond the atmosphere, which became known as Van Allen belts.

Balloons have carried cameras and telescopes to probe various regions of the electromagnetic spectrum, and sent plants and animals to the stratosphere. They have been made of plastic and rubber, and used alone or in fleets. They remain silent and surprisingly stable platforms for science. And for more than science — a series of US research balloons used to study pollution in the 1970s doubled as kinetic art. They are important testing grounds for instruments and techniques that will one day fly in space. "Exploration is the sport of the scientist," Piccard once said. The humble balloon has more than played its part in both, and will continue to do so. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

Scottish science is ready to go it alone

Scientists in restless territories such as Scotland, Quebec and Catalonia should embrace change, **Colin Macilwain** *suggests.*

Referendums are all the rage right now, with British Prime Minister David Cameron pledging that a future Conservative government will hold one on whether the United Kingdom should stay in the European Union. And as Scotland gears up for its own autumn 2014 vote on independence from the United Kingdom, researchers — like everyone else — are starting to contemplate what such a step would mean for their own livelihoods.

Many scientists in Scotland are apprehensive at the prospect of constitutional change. Hugh Pennington, a prominent bacteriologist at the University of Aberdeen, has said that Scotland's researchers should reject independence in the referendum, lest they lose their right to compete for grants from the UK research councils.

I share the opposing view of Stephen Salter, the wave-power pioneer at the University of Edinburgh, who faced Pennington at a recent Royal Society of Chemistry debate on the independence question. Salter says that an independent Scotland would continue to strongly support research, and likens the 'no' argument to the old adage: "Always keep a-hold of nurse, for fear of finding something worse."

Seen from afar, fights for secession can seem parochial and unnecessary. The view from outside is often drenched in superficial sentiment: Canada has its mounted police and low crime; Spain its sunshine and tapas. What on Earth people ask, do those Quebeckers and Catalonians have to complain about?

At least in Scotland's case, outsiders — from continental Europeans tiring of London's endless tantrums over the European Union, to US President Barack Obama, whose grandfather learned all about the British Empire in a detention camp in Kenya in the 1950s — have some inkling of what might be awry in Scotland's 300-year-old union with England.

The university system, together with the armed forces, is one of the few institutions still binding the United Kingdom together. But even at the universities, change is under way. Under the Scotland Act, which restored the Scottish Parliament in 1998, research was one of a handful of powers that were 'reserved' in London, whereas 'the universities' were devolved. In practice, that means that half of the universities' research money now comes through the Scottish government in Edinburgh rather than direct from London — through the university block-grant body, the Scottish Funding Council.

Scotland's higher-education policy has subsequently diverged sharply from its English counterpart: a university education in Scotland remains free, compared with annual tuition fees of up to £9,000 (US$14,000) south of the border.

The UK research councils still operate on a UK-wide basis, however, and it is their loss, in the event of Scottish independence, that Pennington deplores. The seven councils are competently run, but their autonomy — from each other, as well as from Whitehall — has been eroded. Each is chaired by a 'business leader' rather than a scientist, and they routinely irritate researchers with 1980s business-school clap-trap. Unlike whisky, agencies that dispense peer-reviewed grants do not always improve with age: rather, their biases become ingrained. It is feasible to begin afresh: Science Foundation Ireland was started from nothing in 2000 and the European Research Council (ERC) has established a formidable reputation in just six years.

AN INDEPENDENT SCOTLAND WOULD CONTINUE TO STRONGLY SUPPORT RESEARCH.

Mike Russell, the Scottish education minister, is investigating possible approaches for organizing research in an independent Scotland. The options could include contributing to and drawing from the existing research councils. Or Scotland could set up agencies of its own: perhaps one for biomedical research inside the health department and a second for other scientific disciplines. The latter path could better align Scottish university research with Scottish priorities, such as public health, forestry and fisheries, and renewable and offshore energy.

Similar issues arise in other corners of the globe. In Catalonia — which is contemplating its own independence referendum — researchers increasingly look to Brussels, rather than Madrid, for support. The sense that they could go it alone is reinforced by strong performance in Europe-wide competitive peer review: researchers in the province win about three times as many ERC grants per head of population as those in the rest of Spain.

The provincial government in Quebec has steadily assumed greater responsibility for science, although the outlays of its own agency, Research Quebec, are small. Neuroscientist Rémi Quirion was appointed as the province's first chief scientist in 2011.

Last autumn, I asked Ernst Winnacker, former head of both of the German research foundation and the ERC, about the slow speed of university reform in Germany. He spoke wistfully of the strength of the smaller Swiss and Austrian systems. Such sentiments echo those of a housewife in the Scottish town of Kilmarnock, who once told a passing politician that she backed independence because "it's harder to clean a big house than a small house".

The decision on Scotland's future will ride not on blood and thunder, but on such prosaic questions as how best to run science and the universities. Pennington and Salter both happen to be English. But in 2014 they will vote, primarily, on whether the British state or a new creation is better equipped to navigate Scotland through the uncharted waters of the twenty-first century. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/hnubax

Colin Macilwain *writes about science policy from Edinburgh, UK.*
e-mail: cfmworldview@gmail.com

MATERIALS

Trapping water from desert fog

A coated cotton fabric can absorb more than 3 times its weight in water from warm, moist air, and release it again at higher temperatures.

John Xin at the Hong Kong Polytechnic University in China, Catarina Esteves at Eindhoven University of Technology in the Netherlands and their colleagues grafted a temperature-sensitive polymer onto cotton fabric. At 16–25 °C, the polymer chains interact to form a porous, sponge-like network that traps water. At higher temperatures, the material shrinks and releases its moisture in a cycle that is reversible. The material could be useful in dry, desert areas to collect water from morning dew or fog, the authors say.

Adv. Mater. **http://dx.doi.org/10.1002/adma.201204278 (2013)**

S. ABRAMOWICZ/DINOSAUR INST., NHMLAC

PALAEONTOLOGY

Toothy bird had crunchy diet

Many fossil birds have simple teeth, but a fossil found in China has large, grooved teeth and is the first avian fossil to show specialized enamel.

When stomach contents cannot be recovered, palaeontologists look to the teeth for insight into diet and environment. Jingmai O'Connor at the Natural History Museum of Los Angeles County in California and her colleagues describe the fossil of *Sulcavis geeorum* from the early Cretaceous period (145 million to 100 million years ago). The fossil had teeth 1–3 millimetres in length with longitudinal grooves that have never before been seen in a bird.

Whereas small, smooth teeth indicate a herbivorous diet, *S. geeorum* may have used its hard, powerful choppers to crunch creatures with tough exoskeletons, such as insects.

J. Vertebr. Paleontol. **33, 1–12 (2013)**

ANIMAL BEHAVIOUR

Milky Way shows beetles the light

Birds, seals and humans can find their way by the stars — as, it seems, can the dung beetle, using the Milky Way.

Marie Dacke at Lund University in Sweden and her colleagues timed how long the nocturnal dung beetles (*Scarabaeus satyrus*; **pictured**) took to roll their dung balls from the centre of an outdoor arena to its edge. When beetles could see the starlit sky they took less time, and followed straighter paths, than beetles that either had their upward-facing eyes covered or had to navigate on an overcast night. The authors moved their arena into a planetarium, and found that dung beetles exposed to a full starry sky took the same amount of time to exit the arena as those that could see just the Milky Way.

This is the first evidence of an insect navigating using the Milky Way, but it may not be the only animal with this capability, the authors say.

Curr. Biol. **http://dx.doi.org/10.1016/j.cub.2012.12.034 (2013)**

M. BYRNE

ATMOSPHERIC SCIENCE

Predicting storms in East Asia

Forecasting monsoons and tropical storms can be a challenge, but could be improved for East Asia because the variability of a major atmospheric high-pressure system over the western Pacific Ocean seems to be predictable.

Bin Wang and his group at the University of Hawaii at Manoa in Honolulu show that the intensity of the western Pacific Subtropical High is highly correlated with the strength of the summer monsoon and of tropical storm activity. They used climate models to examine the mechanisms that control the system's variability and found that the annual strength and location of the high-pressure system are closely linked to the temperatures of both the central Pacific and Indian oceans.

Understanding the atmosphere-ocean feedbacks that govern atmospheric dynamics could improve the prediction of droughts, floods and storms in the region, the authors suggest.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1214626110 (2013)**

CANCER

Mutations lurk in regulatory regions

Cancer-genome sequencing has yielded a long list of potential cancer-causing mutations, most of which are in genes that code for proteins. But two studies of melanoma genomes have revealed common mutations in a region that regulates gene expression.

Dirk Schadendorf of the University Hospital Essen in Germany, Rajiv Kumar of the German Cancer Research Center in Heidelberg and their colleagues conducted a genetic analysis of 14 members of a family that is prone to the skin cancer. The authors found mutations in a region that regulates the expression of a gene called *TERT*. Another group led by Levi Garraway at the Dana-Farber Cancer Institute in Boston, Massachusetts, found mutations in the same promoter region in 50 out of 70 melanomas. The results suggest that regulatory regions of the genome may be key reservoirs of cancer-causing mutations.

Science **http://dx.doi.org/10.1126/science.1230062; http://dx.doi.org/10.1126/science.1229259 (2013)**

NEUROSCIENCE

Old age, bad sleep, poor memory

The gradual loss of cells in the brain's cortex could be decreasing sleep quality in older adults, leading to poorer long-term memory.

Bryce Mander and Matthew Walker at the University of California, Berkeley, and their group asked healthy adults to memorize a list of words, recall some of them ten minutes later, and recall the rest the next morning. Adults in their late 60s and their 70s performed worse on the test, and showed significant reductions in the slow brain waves that are associated with deep sleep, compared with those around the age of 20. The extent of deep-sleep disruption was related to the degree of memory impairment, and these differences were, in turn, linked with a reduction of grey matter in the medial prefrontal cortex.

The findings suggest that deterioration of this part of the brain diminishes the slow brain waves, which are implicated in memory consolidation, impairing the ability to solidify new memories.

Nature Neurosci. **http://dx.doi.org/10.1038/nn.3324 (2013)**

GENE THERAPY

Gene fix does not prevent cell loss

Gene therapy improves the vision of people with a genetic form of blindness, but does not stop the loss of the light-sensitive photoreceptor cells in the retina.

Childhood blindness as a result of Leber congenital amaurosis (LCA) occurs as a result of photoreceptor dysfunction and degeneration, owing to a mutation in the gene *RPE65*. In addition to fixing the dysfunction, researchers hoped therapy with a working copy of the gene would slow the loss of the cells. But, Artur Cideciyan at the University of Pennsylvania in Philadelphia and his group found that the loss continues. In dogs bearing the LCA mutation, those that were treated at a disease stage approximating to that in humans also showed photoreceptor deterioration.

The results suggest that treatment for this hereditary blindness also needs to address long-term protection of the light-sensing cells.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1218933110 (2013)**

ZOOLOGY

Turtle arrested development

Some reptile species give birth to live young, but turtles have never evolved to do so — perhaps because of low oxygen levels in their egg-laying tubes, or oviducts.

Anthony Rafferty at Monash University in Clayton, Australia, and his group show that oxygen diffused more slowly in secretions from the oviducts of four species of turtle than in saline solution. When the turtles' eggs were incubated at low oxygen levels they stopped developing, whereas those kept at ambient conditions developed normally. The low oxygen levels in the oviducts could explain how turtles are able to store eggs in a state of arrested development until they can lay them on land.

Am. Nat. **http://dx.doi.org/10.1086/668827 (2013)**

COMMUNITY CHOICE

The most viewed papers in science

GENOMICS

Disease genes mutate more

HIGHLY READ on www.cell.com 23 Dec–22 Jan

Genetic mutations occur at random, but where in the genome they occur is non-random. Jonathan Sebat at the University of California, San Diego, Jun Wang at BGI-Shenzhen in China and their group report that some regions of the genome mutate a 100 times more frequently than others, and that genes linked to autism have higher than average mutation rates.

To track emerging mutations, the researchers sequenced the complete genomes of ten sets of identical twins with autism spectrum disorder and their parents. The analysis revealed 'hotspots' where new mutations tend to cluster and showed that mutation rates were associated with certain DNA sequences or with specific aspects of how the DNA is packaged. Disease genes — including those implicated in autism — showed high mutation rates, as did those expressed in the brain.

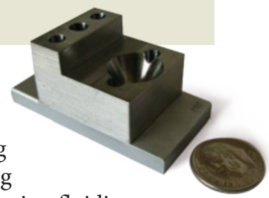
Further study of mutation hotspots could help researchers to identify more genetic risk variants, and to better understand human variability and genome evolution.

Cell 151, **1431–1442 (2012)**

BIOTECHNIQUES

Cell squeezer gets molecules in

The cell membrane, largely impermeable to large molecules, can be breached with needles, electricity and chemicals. But now researchers have devised a less traumatic and more efficient way of delivering molecules into cells, involving squeezing cells in a microfluidic device (**pictured**).



A. SHAREI

A team led by Klavs Jensen and Robert Langer at the Massachusetts Institute of Technology in Cambridge found that passing various cells through the micrometre-wide channels of their device deforms the cells, creating temporary holes in the cell membrane that allow large molecules to pass through. The new approach is 10 to 100 times more efficient than conventional methods at delivering proteins into human skin cells to reprogram them into stem-cell-like cells.

The technique could be used to deliver therapeutic molecules into human cells, the authors say.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1218705110 (2013)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

REED SCHERER

Drilling team reaches Lake Whillans

A US research team drilled through 800-metre-thick ice to reach the subglacial Lake Whillans in Western Antarctica on 28 January. The project is the first to retrieve fully intact samples of liquid water (pictured) and sediment from a subglacial lake, which the team hopes will provide clues to the kind of life that exists in such extreme environments. This is the first time that researchers have probed the water of one of the more than 300 lakes discovered under Antarctica's ice in recent years. See go.nature.com/byj4u8 for more.

RESEARCH

Chimp cut urged

The US National Institutes of Health (NIH) was advised on 22 January to retire most of its 360 chimpanzees to an animal sanctuary and shut down half of its ongoing experiments on the animals. The suggestions were made in a report by independent advisers, who say that about 50 chimpanzees should suffice for future research needs. Francis Collins, director of the NIH, is expected to announce in late March whether the agency will accept the report's recommendations. The report comes in response to advice from the Institute of Medicine in Washington DC, which said in 2011 that most chimp research is unnecessary.

NASA joins Euclid

NASA is joining a €1-billion (US$1.3-billion) European Space Agency mission to explore the 'dark' parts of the Universe. On 24 January the US space agency announced that it would join Euclid, a space telescope that will measure the locations and shapes of some 2 billion distant galaxies. The data will be used to probe dark matter and dark energy. Under the agreement, 40 NASA scientists will join the project and NASA will contribute 20 infrared detectors, valued at around $50 million in total, for one of the instruments on the spacecraft. The mission is scheduled to launch in 2020.

POLICY

Regulator reprieved

Britain's beleaguered regulator of human-embryo research, the Human Fertilisation and Embryology Authority (HFEA), was thrown a lifeline on 25 January. The functions of the HFEA and another regulator — the Human Tissue Authority — were due to be transferred to other bodies as a result of a 2010 government move to cut the numbers of semi-autonomous agencies. After a public consultation rejected the suggestion to close down the two bodies, the Department of Health announced an independent review to assess whether to merge their activities. See go.nature.com/gciolp for more.

Stem-cell reforms

California's US$3-billion stem-cell agency — the California Institute for Regenerative Medicine (CIRM) in San Francisco — is to reform its governance structure to minimize conflicts of interest, the agency's governing board decided on 23 January. The move is in response to an independent review published last month which raised concerns that 13 members of the agency's 29-member board come from research institutions that receive CIRM funds. Board members who represent such institutes will now abstain from votes to approve grants. See go.nature.com/ztij7r for more.

Emissions profits

Airlines that fly to and from Europe may have profited by up to €1.36 billion (US$1.83 billion) last year by raising air fares to cover costs of carbon emissions that they did not actually incur, says a report from CE Delft, a Dutch environmental consultancy group. The European Commission had hoped to

NUMBER CRUNCH

$1.7 m

Total funding for each winner of the Tang Prize, new science prizes announced by Taiwanese billionaire Samuel Yin on 28 January. Starting in 2014, four biennial prizes of US$1.35 million each will be awarded in sustainable development, biopharmaceutical science, law and Chinese studies. Winners can also propose five-year research projects each worth $340,000.

bring intercontinental flights into its 30-nation emissions-trading scheme, and had given airlines some free emissions allowances. But it exempted intercontinental flights from the scheme for 2012, enabling the airlines to achieve windfall profits.

Coffee at risk

Costa Rica has declared a national coffee-growing emergency. The fungus *Hemileia vastatrix*, which causes coffee rust, looks set to wipe out half the nation's 2013–14 harvest in the most affected areas. On 22 January, the government signed an emergency bill to tackle the outbreak. The disease has already attacked coffee crops in South and Central America. See go.nature.com/epwshp and page 587 for more.

Biodiversity panel

The Intergovernmental Platform on Biodiversity and Ecosystem Services — set up in April 2012 to assess the state of the planet's ecosystems — has selected a group of 25 international scientists and ecology experts to safeguard the scientific quality and independence of its work. The appointments were made at a meeting ending on 27 January. Abdul Hamid Zakri, science adviser to the prime minister of Malaysia, was elected as the first chairman of the panel.

PEOPLE

ULRICH BAUMGARTEN/GETTY

Plagiarism inquiry

Germany's science and education minister, Annette Schavan (**pictured**), is being investigated after claims that she plagiarized parts of her PhD thesis in educational science. The University of Düsseldorf announced the inquiry on 23 January after finding that the accusations against Schavan, which were aired last May, are substantive. The minister was awarded her doctorate in 1980 for a study on how conscience develops in people. She denies claims that she quoted the works of others in her thesis without appropriate citation and called on the university to ensure that external experts are involved in the inquiry. See go.nature.com/5phncw for more.

Genomicist dies

David Cox, a pioneering genomicist and senior vice-president at the UK-based drug firm Pfizer, died on 22 January. Cox's research group at Pfizer aimed to find a way to arrange clinical-trial participants on the basis of their genetic make-up. He was also a member of one of the teams that led the Human Genome Project, carried out research on the molecular basis of human genetic disease at Stanford University in California, and was a member of the US National Academy of Sciences.

FUNDING

Global Fund boost

Germany has announced a donation of €1 billion (US$1.3 billion) to the Global Fund to Fight AIDS, Tuberculosis and Malaria for the period 2012–16, of which €600 million is new money. The donation, announced on 24 January, signals support for administrative reforms and staff changes made by the fund last November to address allegations of corruption among its grant recipients. Other states are expected to announce future contributions to the global fund at a fund-raising meeting in September.

Future technologies

The European Commission announced the two winners of its first high-budget competition for future and emerging technologies on 28 January: projects to simulate the human brain and to develop the potential of graphene. The projects should each receive €1 billion (US$1.3 billion) over ten years. See page 585 for more.

COMING UP

7–8 FEBRUARY
London's Royal Society hosts a meeting in Newport Pagnell, UK, to discuss challenges in dealing with storing and indexing massive amounts of research data.
go.nature.com/e2cn8o

7–8 FEBRUARY
In Brussels, European Union states meet again to negotiate the region's budget for 2014–20, including the amount apportioned to research, for which around €80 billion (US$100 billion) has been proposed. Talks broke down last year.
go.nature.com/2kq2ua

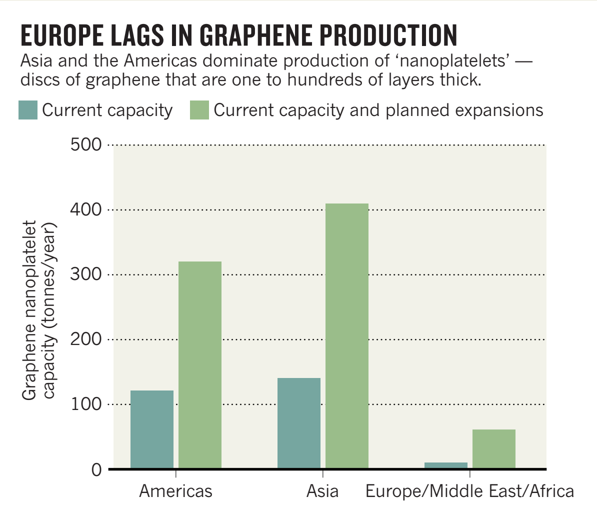
EVENTS

MIT hacked again

The website of the Massachusetts Institute of Technology (MIT) in Cambridge was hacked on 22 January for the second time in a week. The attacks are a protest at the suicide of Aaron Swartz, an Internet activist who killed himself earlier this month. Swartz had been charged with using MIT data networks illegally, by downloading millions of academic articles from JSTOR, a scholarly archive; he faced up to 35 years in prison and heavy fines. MIT has asked one of its computer scientists, Hal Abelson, to review the university's conduct in the affair. Abelson expects to complete a report in the next few weeks.

TREND WATCH

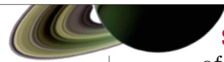
Graphene research in Europe has a funding boost (see page 585), but the commercial action is hotter elsewhere. Multilayered flakes or discs of graphene — 'nanoplatelets' — which may find use in adding strength and conductivity to composites and coatings, are mainly produced in the Americas and Asia, according to analysts Lux Research in Boston, Massachusetts. Planned capacity expansion in China could see supply of the platelets outstrip demand, Lux analyst Ross Kozarsky adds.



SOURCE: LUX RESEARCH

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS



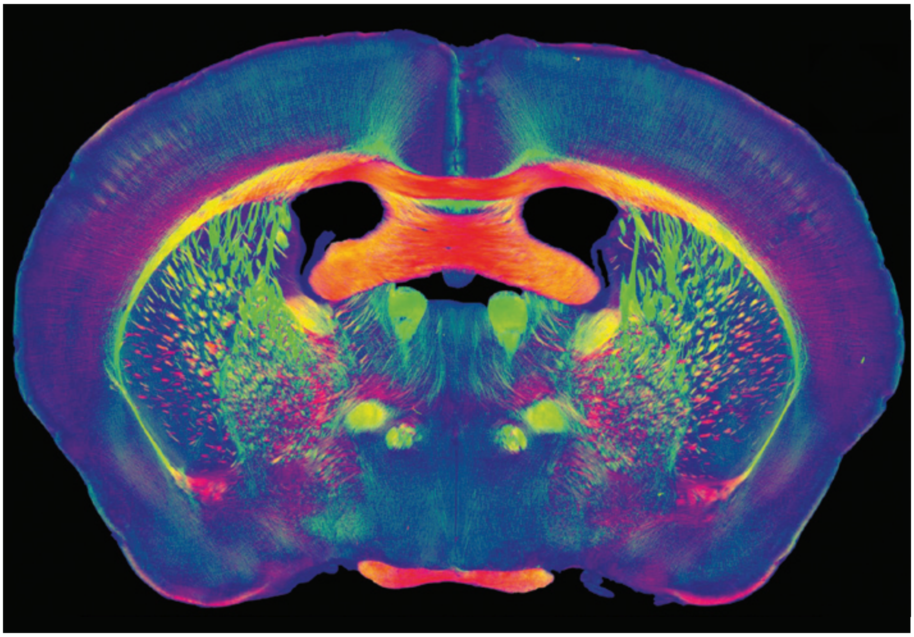
AGRICULTURE Coffee rust wilts harvests but spurs research effort **p.587**

TECHNOLOGY Magnetic circuits could be rewired on the fly **p.589**

CLIMATE Obama gives greenhouse policy a second wind **p.590**

SPACE Are the glories of the outer Solar System a passing show? **p.592**

FZJ



The Human Brain Project will also study the mouse brain to build its prizewinning simulation.

FUNDING

Research prize boost for Europe

Graphene and virtual brain win billion-euro competition.

BY ALISON ABBOTT AND QUIRIN SCHIERMEIER

Two of the biggest awards ever made for research have gone to boosting studies of the wonder material graphene and an elaborate simulation of the brain. The winners of the European Commission's two-year Future and Emerging Technologies 'flagship' competition, announced on 28 January, will receive €500 million (US$670 million) each for their planned work, which the commission hopes will help to improve the lives, health and prosperity of millions of Europeans.

The Human Brain Project, a supercomputer simulation of the human brain conceived and led by neuroscientist Henry Markram at the Swiss Federal Insitute of Technology in Lausanne, scooped one of the prizes. The other winning team, led by Jari Kinaret at Chalmers University of Technology in Gothenburg, Sweden, hopes to develop the potential of graphene — an ultrathin, flexible, electrically conducting form of carbon — in applications such as personal-communication technologies, energy storage and sensors.

The size of the awards — matching funds raised by the participants are expected to bring each project's budget up to €1 billion over ten years — have some researchers worrying that the flagship programme may draw resources from other research. And both winners have already faced criticism.

Many neuroscientists have argued, for example, that the Human Brain Project's approach to modelling the brain is too cumbersome to succeed (see *Nature* **482,** 456–458; 2012). Markram is unfazed. He explains that the project will have three main thrusts. One will be to study the structure of the mouse brain, from the molecular to the cellular scale and up. Another will generate similar human data. A third will try to identify the brain wiring associated with particular behaviours. The long-term goals, Markram says, include improved diagnosis and treatment of brain diseases, and brain-inspired technology.

"It's a very bold project," says Mark Fishman, president of the Novartis Institutes for BioMedical Research in Cambridge, Massachusetts, adding that it will "no doubt spawn unexpected new research directions, probably to help develop supercomputing and medical robotics". No one knows exactly what data will be needed to simulate the human brain, he says — "the Human Brain Project will help us find out".

The graphene project faces the criticism that the potential of the one-atom-thick sheets of carbon, first reported by Andre Geim and Konstantin Novoselov of the University of Manchester, UK, in 2004, may be overhyped. The material's extraordinary line-up of properties — transparency, electrical conductivity, flexibility and strength — has wowed industry and academia alike, leading to visions of cheap solar cells and large-screen mobile phones that can be rolled up to pocket size. It also won Geim and Novoselov the 2010 Nobel Prize in Physics.

But analysts caution that graphene's properties are no guarantee of commercial success. "Major challenges, such as high costs, processing issues and competing materials, loom large," cautions Ross Kozarsky, who leads the advanced-materials team at business analysts Lux Research in Boston, Massachusetts.

Kinaret thinks that the steady, long-term funding for graphene research under the flagship project should help to address those problems, and could allow Europe to pull ahead of Asia in commercializing the material.

But whether either project will receive all of the promised funding is unclear. The European Union's Seventh Framework Programme of research, which ends in December, has provided €108 million to support the 'ramp-up' phase of the two winning projects for their first 30 months. ▸

➲ NATURE.COM
See the *Nature* Outlook for more on graphene:
go.nature.com/otkrsb

Its 2014–20 successor, Horizon 2020, will support the second phase. But the Horizon 2020 budget is likely to fall well short of the €77.6 billion proposed by the commission, and some observers fear that support earmarked for the flagships may be scaled back as a result.

Thousands of scientists across Europe worked intensively on developing the diverse project bids that were submitted to the competition — from computerized personal medicine to perceptive robots that respond to human needs. Some participants complain that the competition's goalposts seemed to shift during the selection process. At first, they claim, the commission stressed that winning projects would be chosen mainly for their scientific excellence. "But it became clear that impact for economic growth and for consumers was becoming more important — understandable in the economic climate," says Kinaret.

However, Wolfgang Boch, head of the commission's flagship unit, says that a panel of 25 experts from science and industry eventually chose the two winners on the basis of the published criteria, and that scientific excellence counted for 50% of the final ranking.

Even losers say they benefited from the competition. When the interim rankings were published last July, FuturICT was tipped to win. That project aims to model human activities and their impact on global political stability, the environment and financial markets. Its coordinator, Dirk Helbing, a physicist-turned-sociologist at the Swiss Federal Institute of Technology in Zurich, says he is disappointed, but that the interdisciplinary community the project created "will stay alive and active".

"We know that covert FuturICT-like projects are being started in other parts of the world," he says. "That makes it even more important to continue our open, transparent and participatory project." ■

FUNDING

UK research councils could face mergers

Wide-ranging review edges towards single funding pot.

BY GEOFF BRUMFIEL

A government review that quietly began earlier this month could lead to major changes at the agencies charged with distributing much of the United Kingdom's scientific funding.

Possible changes to improve efficiency include bringing the roughly £3-billion (US$4.7-billion) annual spend of all seven research councils into a single pot — potentially resulting in a body that would look rather like the US National Science Foundation (NSF). But observers fear that such a shake-up could bring years of chaos and disrupt the links between funders and the communities they serve.

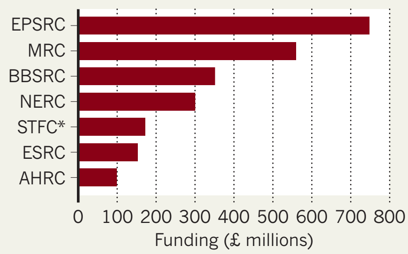
At a minimum, says David Price, vice-provost for research at University College London, the recommendations "will have wide implications for the research sector". But he adds, "I don't think it's widely known that this is going on".

Conducted at the request of the powerful Cabinet Office, the review is designed to provide a robust challenge to the continuing need for the councils. It will also examine their current structure and may recommend reducing their number or consolidating them into a single grant-funding body.

The review, expected to be completed by the summer, is part of a broader examination of independent government bodies by Francis Maude, the chief cabinet minister. Last August, Maude announced that the government had already abolished 100 quasi-governmental organizations, and promised more cuts to come. The research councils operate independently of their parent Department for Business Innovation and Skills (BIS), and as such, they are subject to the review. At its most extreme, the review committee could recommend that the councils are brought under the direct control of the BIS, which funds them, or be spun off as totally independent, charity-like bodies — although presumably still receiving government cash.

MAINTAIN OR MERGE?

Grant funding allocations for the seven UK research councils in 2013–14.



EPSRC, Engineering and Physical Sciences Research Council; MRC, Medical Research Council; BBSRC, Biotechnology and Biological Sciences Research Council; NERC, Natural Environment Research Council; STFC, Science and Technology Facilities Council; ESRC, Economic and Social Research Council; AHRC, Arts and Humanities Research Council.
*Budget for core programme only.

David Willetts, the government's minister for universities and science, doubts that such radical revisions are in store. "My view is that the model works pretty well, and I would be surprised if the review reached radical conclusions," he says. "But if there are lessons about how [the councils] can raise their game, we'll look at it."

➲ NATURE.COM
For more on UK research council controversies, see: go.nature.com/fryeo9

Past reviews of the councils have led to big changes, however. A 2001 review spawned an overarching body called Research Councils UK, which helps to coordinate the activities of the councils. A follow-up in 2004 led to the creation of a central system for the councils' human resources, information technology, finance, grants and recruitment.

Despite this recent consolidation, the councils remain largely independent bodies, with their own chief executives, advisory boards and budgets. "They represent very different functions and communities," says Luke Georghiou, vice-president for research and innovation at the University of Manchester, who participated in the 2001 review.

That could change with the latest review, which is being led by Ceri Smith, director of labour markets at the BIS, and an outsider to the academic community. Smith is believed to be considering various options, including consolidating several of the councils or appointing a single official to oversee the budgets of all of them (see 'Maintain or merge?'). Such changes might reduce administrative costs.

Pulling the research councils' budgets into a single pot might effectively create a single council similar to the NSF. That would be a mistake, says Georghiou. Unlike the NSF, which functions mainly to award and disburse grants, the councils have a diversity of obligations to the scientists they serve, including the running of institutes and facilities.

Research administrators also say that merging several councils could be especially unsettling at a time of tight budgets. Price serves on the board of the Science and Technology Facilities Council, which was formed out of a merger of two smaller councils in 2007. Budget cuts and administrative problems dogged the new council for years after the merger, he says. "The scars have just about healed now, but it took loads of time."

If yet more councils are fused, says Georghiou, "we could face three years of disruption at a time when we have to make maximum strategic use of what there is". ■

AGRICULTURE

Coffee rust regains foothold

Researchers marshal technology in bid to thwart fungal outbreak in Central America.

HECTOR RETAMAL/AFP/GETTY

BY DANIEL CRESSEY

Coffee growers are worried that a fungal outbreak will affect the next harvest of coffee berries.

Where there is coffee, there is 'coffee rust'. But the long stalemate between growers and the fungus behind the devastating disease has broken — with the fungus taking the advantage. As one of the most severe outbreaks ever rages through Central America, researchers are reaching for the latest tools in an effort to combat the pest, from sequencing its genome to cross-breeding coffee plants with resistant strains.

Caused by the fungus *Hemileia vastatrix*, coffee rust generally does not kill plants, but the Institute of Coffee of Costa Rica estimates that the latest outbreak may halve the 2013–14 harvest in the worst affected areas of the nation. This outbreak is "the worst we've seen in Central America and Mexico since the rust arrived" in the region more than 40 years ago, says John Vandermeer, an ecologist at the University of Michigan in Ann Arbor, who has received "reports of devastation in Nicaragua, El Salvador and Mexico".

At his research plot in Mexico, Vandermeer says that the situation is so bad that the leaves are simply dropping off the plants. More than 60% of the trees have at least 80% defoliation, and 30% have no leaves at all.

On 22 January, Costa Rica enacted emergency legislation to speed up the flow of government money towards fighting the fungus. Other nations are also stepping up the fight. Last week, the Nicaraguan government reportedly declared that it would include coffee rust on a list of special research projects designed to safeguard the country's agriculture.

The fungus first emerged as a significant problem by 1869 in Ceylon — now Sri Lanka — before spreading around the world. Stuart McCook, a historian at the University of Guelph in Canada who studies the rust, says that the wet weather in some areas of Ceylon was ideal for the spread of the fungus, and more than 90% of coffee crops were wiped out in those regions. Faced with an economic catastrophe, the country abandoned coffee for the tea it is associated with today. The disease is so universal that it "is not going to be eradicated; or the only way to eradicate the disease in practice is to eradicate all of the coffee", says McCook.

By 1970, the fungus had been detected in Brazil, and severe outbreaks were seen in Costa Rica in 1989 and Nicaragua in 1995, says Jacques Avelino, a plant pathologist at Costa Rica's Tropical Agricultural Research and Higher Education Center, based in San José.

But changes to management practices had brought the disease mostly under control. "Coffee rust was considered a solved problem by most of the coffee growers and coffee institutes of the region", says Avelino. "People didn't fear the disease." The outbreak may have taken hold because of patchy use and effectiveness of fungicides.

And in Africa, Noah Phiri, a plant pathologist working in Nairobi for the not-for-profit development organization CABI, says that rust has been causing ever-greater problems, although in Kenya, varieties resistant to the rust have held it at bay.

Colombia could be the closest to a solution. Marco Aurelio Cristancho, a researcher at Cenicafé, the National Centre for the Investigation of Coffee in Chinchiná, says that the government has supported research into developing resistant strains of coffee through cross-breeding. The introduction of resistant strains, together with improved weather monitoring to help predict rust outbreaks, has meant that fewer than 10% of plants now need to be treated with fungicide, down from 60% four years ago, Cristancho says. The government has also supported work on the genetics of both the fungus and the plant.

Research programmes have started in other countries, too. At the Federal Rural University of Rio de Janeiro in Brazil, Valdir Diola is working to isolate resistance genes in coffee and to find molecular markers that distinguish between different strains of the pathogen and that could be used to develop tailored strategies for its control. And in the United Kingdom, Harry Evans is working on the genome of *H. vastatrix* at CABI in Egham. In Nairobi, Phiri is using money from the intergovernmental agency the Common Fund for Commodities, as well as from Kenya, India, Rwanda, Uganda and Zimbabwe, to screen for resistant coffee plants and to analyse varieties of the pathogen.

"Scientists need to continuously develop resistant varieties in order to keep coffee leaf rust disease at bay," Phiri says. "Governments in coffee-growing countries need to take coffee research as a priority and provide necessary resources."

Cristancho says that other nations need to adopt an integrated approach similar to that of Colombia. "Unfortunately this effort is not mirrored in other regions of the world, where it is required to provide local solutions to the epidemics," he says. ■

NIGEL CATTLIN/FLPA

***Hemileia vastatrix* 'rusts' the leaves of coffee plants.**

➲ NATURE.COM
Read about efforts to produce caffeine-free coffee strains:
go.nature.com/m5e66i

RESEARCH

Funding agencies urged to check for duplicate grants

Nature probe reveals lack of oversight of researchers who win two grants for similar projects.

BY EUGENIE SAMUEL REICH AND CONOR L. MYHRVOLD

When neuroscientist Steven McIntire of the University of California, San Francisco, submitted a five-year, US$1.6-million grant application to the US National Institutes of Health (NIH) in November 2001, he did not mention that just five months earlier, the US Army had awarded him $1.2 million for a project with strikingly similar scientific aims. Both grants supported a search for genes that affect responses to ethanol in the worm *Caenorhabditis elegans*, which is used as a model organism to understand the effects of alcohol in humans.

McIntire, who is no longer in research but sees patients at Stanford Hospital in California, says that the two grants paid for different research: the army funds were used to look for ethanol-resistance genes, whereas the NIH's cash was spent on pinning down ethanol-hypersensitivity genes. There is no implication that McIntire or any of the other researchers connected to the cases in this news story committed any wrongdoing.

But the NIH remained unaware of the army grant, and its similarity to the NIH application, throughout peer review and initial evaluation of McIntire's grant. It ultimately learned of the similarity from McIntire himself. Given that the agency wants to avoid awarding duplicate grants, the case raises questions about how effectively funders screen applications for overlap.

"The agencies are overwhelmed, and checking grants at other agencies is something that doesn't exist," says bioinformatician Harold Garner at the Virginia Polytechnic Institute and State University in Blacksburg. In a Comment article in this week's *Nature* (see page 599), Garner and his colleagues estimate that nearly $70 million in overlapping funds may have been awarded over the past decade — money that could potentially have been spent on more original research.

"The average grant is about $450,000. A couple of days of labour to avoid overlap should be worth that."

They came to that figure after reviewing US grant applications in publicly

IVAN MOROZOV, VIRGINIA BIOINFORMATICS INST.

Harold Garner has used text-similarity software to identify overlap in grant applications.

accessible databases. A computerized search for duplicated text turned up 1,300 applications with potential overlap, from some 850,000 grant applications. After manually reviewing those cases, Garner's team pulled out 167 pairs that were very similar.

Because they did not have access to the full grant files, which would have allowed them to do a more thorough assessment, the team has opted not to identify those grants. But they did provide the data to *Nature*'s news team, which subsequently obtained documentation on 22 pairs of very similar grants through the US Freedom of Information Act (FOIA). After an examination of these files, about half of the potential duplications seemed to warrant further investigation (see 'Doubling up'). This confirms that a computerized search for duplicated text "is a method to find [overlapping] grants that need to have adjustments", says Garner. The agency could then respond by reducing funding or insisting on a change in research goals.

The potential for overlap has risen since electronic preparation and submission of grant applications became the norm, says Karen Markin, who supervises grant-raising in her role as the director of research development at the University of Rhode Island in Providence. She says that it is now easier than ever for scientists to cut and paste from one document to the next.

Meanwhile, the US Congress has begun to take more interest in the issue: in an audit last year (www.gao.gov/duplication), its Government Accountability Office concluded that both the NIH and the US Department of Defense should do more to avoid duplication in their funding of health research.

Nature's review of the grant files suggests that US agencies often fail to document whether they have looked for similarities in grant proposals, or what actions they took when possible overlaps occurred. In response, the agencies say that they have a number of measures in place, including requiring researchers and their institutions to declare duplicate submissions, as well as a peer-review process that sometimes catches overlap before funding is awarded.

In McIntire's case, the potential for overlap did not come to light until well after his application had been peer reviewed, says the NIH, although before the award was made. The US Army, by contrast, learned of his subsequent NIH grant only in 2003, when reviewing McIntire's progress report, according to comments and grant files supplied by both agencies to *Nature* under the FOIA. In a disclosure to the NIH, McIntire wrote that although the projects had initially overlapped, they later diverged sufficiently that "there may be additional synergy between the two grants, but no scientific or financial overlap".

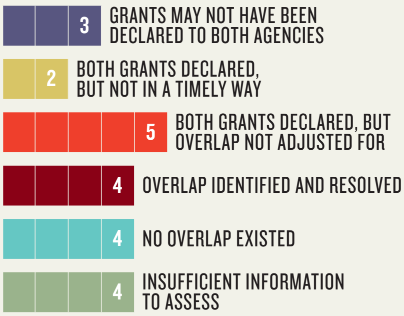
REGULATION LABYRINTH

Determining how much overlap is acceptable is not easy given the thicket of rules that researchers must negotiate. The NIH, for example, prohibits any scientific overlap in the projects it funds. The US National Science Foundation (NSF), by contrast, requests that researchers alert it to the submission of identical proposals to different agencies on an application cover-sheet, and that they use progress reports to inform it of changes in the grants they hold. The US Department of Defense does not always require this type of notification, but does place responsibility on researchers to ensure that no overlap exists. Not only do regulations vary between agencies, but practice on the ground is also inconsistent, says research ethicist C. K. Gunsalus of the

➲ NATURE.COM
For responses from funding agencies on grant overlaps, see:
go.nature.com/asbxgv

DOUBLING UP

A review of agency documents for 22 grant pairs flagged up by an automated search for duplicated text suggested that about half warranted closer scrutiny.



SOURCES: NIH, NSF, DOD, US ARMY

University of Illinois at Urbana-Champaign. "It's a morass," she says.

That leaves plenty of room for confusion. For example, when medical researcher Allen Gao of the University of California, Davis, won an army grant to study androgen-receptor signalling in prostate cancer, officials there worked with him to change his goals so that his application would not overlap with an NIH grant that he had been awarded in 2001. But when Gao disclosed the army grant to the NIH in a 2002 progress report, the agency's officials began what they termed "extensive discussions" with Gao. They concluded that his two grants still overlapped, and reduced the NIH grant by $75,000. "I believed that the issue had been resolved," says Gao.

Researchers say that they are eager for clarity about the limits of acceptable behaviour when chasing funding. The NIH explains that it is acceptable for researchers to submit similar requests to different agencies without disclosing other grants, because this information is required only at what is called the 'just-in-time' stage, before an award is finalized.

Some researchers already follow that principle. In one case identified by Garner's search, Michael Zuscik, a medical researcher at the University of Rochester in New York, submitted identical proposals to the army and the NIH. But adjustments he made to the NIH proposal in response to a reviewer's scientific critique removed overlap with the army grant, he says. "This approach to securing support for research is a common method — submit the aims to more than one appropriate funding agency in the hope one will 'hit'."

Zuscik used the army money to test the effect of nicotine on fracture-healing in mice, and the NIH funds to test the effect of cigarette smoke on the same process — different science that nevertheless required some of the same control experiments. Zuscik says that the repetition was required to ensure scientific rigour. Although grant files show that Zuscik alerted the NIH to the army grant, the army has told *Nature* that it had not been aware of the NIH funding, and is now researching both awards to see whether they overlap.

Michael Emch, a geographer at the University of North Carolina at Chapel Hill, argues that there will always be at least some intellectual overlap between different projects run by the same researcher. Two of his grants were picked out by Garner's search because they had the same title and similar abstracts. A review of the full grant applications shows that the hypotheses and much of the text describing their methodologies is also identical. But Emch says that he did not charge both agencies for the same expenses, such as labour and lab equipment. Emch has an extremely broad research programme, and the NIH money was used to apply general medical-geography methods that had been developed with funds from the NSF to study cholera, he says.

Gunsalus points out that such grants may overlap for practical, as well as intellectual, reasons. Researchers who have large projects may carve out different lines of inquiry within them but submit similar grant applications for each one; or they might use seed funding from one agency to start a project, then try to raise additional funds from another.

Garner insists that agencies need more-consistent regulations and definitions of overlap. He also advocates for a central grant database that flags duplicated text automatically — although a manual review would still be required to pin down whether overlap exists. "The average grant is about $450,000. A couple of days of labour to avoid overlap should be worth that," he says. ■

ELECTRONICS

Magnetic logic makes for mutable chips

Alternative transistor relies on exotic semiconductor.

BY GEOFF BRUMFIEL

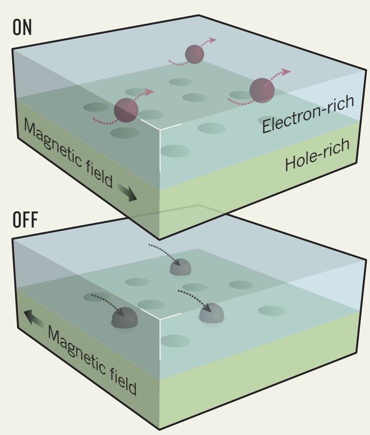
Software can transform a computer from a word processor to a number cruncher to a video telephone. But the underlying hardware is unchanged. Now, a type of transistor that can be switched with magnetism instead of electricity could make circuitry malleable too, leading to more efficient and reliable gadgets, from smart phones to satellites.

Transistors, the simple switches at the heart of all modern electronics, generally use a tiny voltage to toggle between 'on' and 'off'. The voltage approach is highly reliable and easy to miniaturize, but has its disadvantages. First, keeping the voltage on requires power, which drives up the energy consumption of the microchip. Second, transistors must be hard-wired into the chips and can't be reconfigured, which means computers need dedicated circuitry for all their functions.

A research group based at the Korea Institute of Science and Technology (KIST) in Seoul, South Korea, has developed a circuit that may get around these problems. The device, described in a paper published on *Nature*'s website on 30 January, uses magnetism to control the flow of electrons across a minuscule bridge of the semiconducting material indium antimonide (S. Joo *et al. Nature* http://dx.doi.org/10.1038/nature11817; 2013). It is "a new and interesting twist on how to implement a logic gate", says Gian Salis, a physicist at IBM's Zurich Research Laboratory in Switzerland.

MAGNETIC LOCK

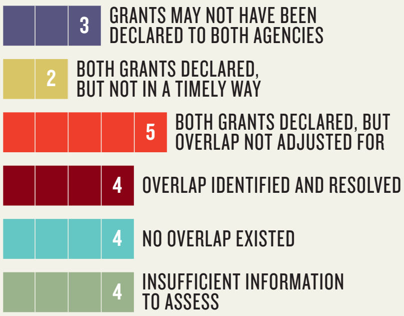
In a circuit made of the semiconductor indium antimonide, a magnetic field can lift electrons over positively charged holes, switching the device on — or deflect them into the holes, turning it off.



The bridge has two layers: a lower deck with an excess of positively charged holes and an upper deck filled predominantly with negatively charged electrons. Thanks to the unusual electronic properties of the indium antimonide, the researchers can control the flow of electrons across the bridge using a ▶

DOUBLING UP

A review of agency documents for 22 grant pairs flagged up by an automated search for duplicated text suggested that about half warranted closer scrutiny.



SOURCES: NIH, NSF, DOD, US ARMY

University of Illinois at Urbana-Champaign. "It's a morass," she says.

That leaves plenty of room for confusion. For example, when medical researcher Allen Gao of the University of California, Davis, won an army grant to study androgen-receptor signalling in prostate cancer, officials there worked with him to change his goals so that his application would not overlap with an NIH grant that he had been awarded in 2001. But when Gao disclosed the army grant to the NIH in a 2002 progress report, the agency's officials began what they termed "extensive discussions" with Gao. They concluded that his two grants still overlapped, and reduced the NIH grant by $75,000. "I believed that the issue had been resolved," says Gao.

Researchers say that they are eager for clarity about the limits of acceptable behaviour when chasing funding. The NIH explains that it is acceptable for researchers to submit similar requests to different agencies without disclosing other grants, because this information is required only at what is called the 'just-in-time' stage, before an award is finalized.

Some researchers already follow that principle. In one case identified by Garner's search, Michael Zuscik, a medical researcher at the University of Rochester in New York, submitted identical proposals to the army and the NIH. But adjustments he made to the NIH proposal in response to a reviewer's scientific critique removed overlap with the army grant, he says. "This approach to securing support for research is a common method — submit the aims to more than one appropriate funding agency in the hope one will 'hit'."

Zuscik used the army money to test the effect of nicotine on fracture-healing in mice, and the NIH funds to test the effect of cigarette smoke on the same process — different science that nevertheless required some of the same control experiments. Zuscik says that the repetition was required to ensure scientific rigour. Although grant files show that Zuscik alerted the NIH to the army grant, the army has told *Nature* that it had not been aware of the NIH funding, and is now researching both awards to see whether they overlap.

Michael Emch, a geographer at the University of North Carolina at Chapel Hill, argues that there will always be at least some intellectual overlap between different projects run by the same researcher. Two of his grants were picked out by Garner's search because they had the same title and similar abstracts. A review of the full grant applications shows that the hypotheses and much of the text describing their methodologies is also identical. But Emch says that he did not charge both agencies for the same expenses, such as labour and lab equipment. Emch has an extremely broad research programme, and the NIH money was used to apply general medical-geography methods that had been developed with funds from the NSF to study cholera, he says.

Gunsalus points out that such grants may overlap for practical, as well as intellectual, reasons. Researchers who have large projects may carve out different lines of inquiry within them but submit similar grant applications for each one; or they might use seed funding from one agency to start a project, then try to raise additional funds from another.

Garner insists that agencies need more-consistent regulations and definitions of overlap. He also advocates for a central grant database that flags duplicated text automatically — although a manual review would still be required to pin down whether overlap exists. "The average grant is about $450,000. A couple of days of labour to avoid overlap should be worth that," he says. ■

ELECTRONICS

Magnetic logic makes for mutable chips

Alternative transistor relies on exotic semiconductor.

BY GEOFF BRUMFIEL

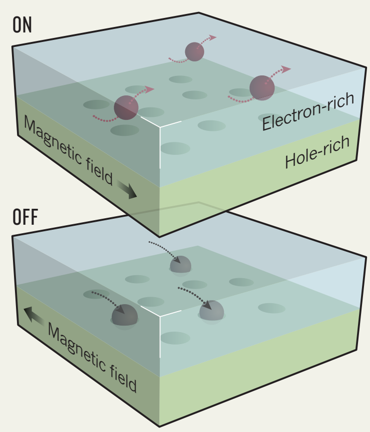
Software can transform a computer from a word processor to a number cruncher to a video telephone. But the underlying hardware is unchanged. Now, a type of transistor that can be switched with magnetism instead of electricity could make circuitry malleable too, leading to more efficient and reliable gadgets, from smart phones to satellites.

Transistors, the simple switches at the heart of all modern electronics, generally use a tiny voltage to toggle between 'on' and 'off'. The voltage approach is highly reliable and easy to miniaturize, but has its disadvantages. First, keeping the voltage on requires power, which drives up the energy consumption of the microchip. Second, transistors must be hard-wired into the chips and can't be reconfigured, which means computers need dedicated circuitry for all their functions.

A research group based at the Korea Institute of Science and Technology (KIST) in Seoul, South Korea, has developed a circuit that may get around these problems. The device, described in a paper published on *Nature*'s website on 30 January, uses magnetism to control the flow of electrons across a minuscule bridge of the semiconducting material indium antimonide (S. Joo *et al. Nature* http://dx.doi.org/10.1038/nature11817; 2013). It is "a new and interesting twist on how to implement a logic gate", says Gian Salis, a physicist at IBM's Zurich Research Laboratory in Switzerland.

MAGNETIC LOCK

In a circuit made of the semiconductor indium antimonide, a magnetic field can lift electrons over positively charged holes, switching the device on — or deflect them into the holes, turning it off.



The bridge has two layers: a lower deck with an excess of positively charged holes and an upper deck filled predominantly with negatively charged electrons. Thanks to the unusual electronic properties of the indium antimonide, the researchers can control the flow of electrons across the bridge using a ▸

perpendicular magnetic field. When they set the field in one direction, electrons are steered away from the positive bottom deck and flow freely. When the magnetic field is flipped, the electrons crash into the lower deck and recombine with the holes — effectively turning the switch off (see 'Magnetic lock').

The ability of a magnetic logic gate to hold the switch on or off without a voltage "could lead to great reduction of energy consumption", says study co-author Jin Dong Song, a physicist at KIST. Even more impressively, the magnetic switches "can be handled like software", he says, by simply flipping the field to enable or disable a circuit. Thus a mobile phone could, for example, reprogram a bit of its microcircuitry to process video while its user watched a clip on YouTube, then switch the chip back to signal processing to take a phone call. This could greatly reduce the volume of circuitry needed inside the phone.

Such reconfigurable logic could be invaluable in satellites, adds Mark Johnson of the Naval Research Laboratory in Washington DC, a co-author of the paper. If part of a chip failed in orbit, another sector could simply be reprogrammed to take over. "You've healed the circuit and you've done it from Earth," he says.

To really catch on, however, the magnetic logic would have to be integrated with existing silicon-based technologies. That may not be easy. For one thing, indium antimonide, the semiconductor crucial to the circuits, doesn't lend itself well to manufacturing processes used to make modern electronics, according to Junichi Murota, a researcher working with nanoelectronics at Tohoku University in Japan. But Johnson says that it may eventually be possible to build similar bridges with silicon.

Integrating the miniature magnets needed to control the devices into a normal chip wouldn't be easy either. Companies should be able to solve these challenges, but only if they decide the devices are worthwhile, says Salis. At the moment, he adds, it is not clear whether the devices will perform well at the sizes needed for a practical chip — much smaller than the micrometre dimensions of the prototypes.

But Johnson notes that magnetism is already catching on in circuit design: some advanced devices are beginning to use a magnetic version of random access memory, a type of memory that has historically been built only with conventional transistors. "I think a shift is already under way," he says. ■

ZUMA/REX FEATURES

US President Barack Obama reinforced environment promises in his second inaugural address.

ENVIRONMENT

Obama rekindles climate hopes

President will use regulations to sidestep stalled Congress.

BY JEFF TOLLEFSON

Throughout his re-election campaign, US President Barack Obama rarely said the words 'climate change'. But in his second inaugural address, on 21 January, Obama renewed a commitment to address global warming, citing both moral and economic imperatives. To fail, he said, "would betray our children and future generations".

The 2010 demise of a climate bill that would have enacted a cap-and-trade system to limit greenhouse-gas emissions remains one of the key failures of Obama's first term. With a divided Congress still standing in the way of legislation, the administration is likely to rely on its own power to impose new regulations, once Obama has replaced the retiring heads of three agencies key to the climate agenda (see 'Climate team change').

As proof of what is possible, Obama can point to a welcome, if unexpected, reduction in US greenhouse-gas emissions during his first term. The decline is in part a result of the economic slowdown and a shift in electricity production from coal to natural gas, which has become cheap and plentiful in recent years. But policies have helped. These include federal greenhouse-gas standards for vehicles, and the introduction by more than half of the states of significant energy and climate initiatives that could deliver further

MORE ONLINE

TOP STORY

JACK ATLEY/BLOOMBERG/GETTY

Australian research agency raises ongoing questions on alleged bullying go.nature.com/sqvzlk

MORE NEWS

- Small oil spills may actually be bigger than originally thought go.nature.com/a64awg
- Ageing causes poor sleep and impairs memory go.nature.com/hppcay
- Aphrodisiac craze endangers Himalayan caterpillar go.nature.com/yegirz

SLIDESHOW

NATURAL HISTORY MUSEUM

Letters of Alfred Russel Wallace go online go.nature.com/jepahh

perpendicular magnetic field. When they set the field in one direction, electrons are steered away from the positive bottom deck and flow freely. When the magnetic field is flipped, the electrons crash into the lower deck and recombine with the holes — effectively turning the switch off (see 'Magnetic lock').

The ability of a magnetic logic gate to hold the switch on or off without a voltage "could lead to great reduction of energy consumption", says study co-author Jin Dong Song, a physicist at KIST. Even more impressively, the magnetic switches "can be handled like software", he says, by simply flipping the field to enable or disable a circuit. Thus a mobile phone could, for example, reprogram a bit of its microcircuitry to process video while its user watched a clip on YouTube, then switch the chip back to signal processing to take a phone call. This could greatly reduce the volume of circuitry needed inside the phone.

Such reconfigurable logic could be invaluable in satellites, adds Mark Johnson of the Naval Research Laboratory in Washington DC, a co-author of the paper. If part of a chip failed in orbit, another sector could simply be reprogrammed to take over. "You've healed the circuit and you've done it from Earth," he says.

To really catch on, however, the magnetic logic would have to be integrated with existing silicon-based technologies. That may not be easy. For one thing, indium antimonide, the semiconductor crucial to the circuits, doesn't lend itself well to manufacturing processes used to make modern electronics, according to Junichi Murota, a researcher working with nanoelectronics at Tohoku University in Japan. But Johnson says that it may eventually be possible to build similar bridges with silicon.

Integrating the miniature magnets needed to control the devices into a normal chip wouldn't be easy either. Companies should be able to solve these challenges, but only if they decide the devices are worthwhile, says Salis. At the moment, he adds, it is not clear whether the devices will perform well at the sizes needed for a practical chip — much smaller than the micrometre dimensions of the prototypes.

But Johnson notes that magnetism is already catching on in circuit design: some advanced devices are beginning to use a magnetic version of random access memory, a type of memory that has historically been built only with conventional transistors. "I think a shift is already under way," he says. ■

ZUMA/REX FEATURES

US President Barack Obama reinforced environment promises in his second inaugural address.

ENVIRONMENT

Obama rekindles climate hopes

President will use regulations to sidestep stalled Congress.

BY JEFF TOLLEFSON

Throughout his re-election campaign, US President Barack Obama rarely said the words 'climate change'. But in his second inaugural address, on 21 January, Obama renewed a commitment to address global warming, citing both moral and economic imperatives. To fail, he said, "would betray our children and future generations".

The 2010 demise of a climate bill that would have enacted a cap-and-trade system to limit greenhouse-gas emissions remains one of the key failures of Obama's first term. With a divided Congress still standing in the way of legislation, the administration is likely to rely on its own power to impose new regulations, once Obama has replaced the retiring heads of three agencies key to the climate agenda (see 'Climate team change').

As proof of what is possible, Obama can point to a welcome, if unexpected, reduction in US greenhouse-gas emissions during his first term. The decline is in part a result of the economic slowdown and a shift in electricity production from coal to natural gas, which has become cheap and plentiful in recent years. But policies have helped. These include federal greenhouse-gas standards for vehicles, and the introduction by more than half of the states of significant energy and climate initiatives that could deliver further

MORE ONLINE

TOP STORY

JACK ATLEY/BLOOMBERG/GETTY

Australian research agency raises ongoing questions on alleged bullying go.nature.com/sqvzlk

MORE NEWS

- Small oil spills may actually be bigger than originally thought go.nature.com/a64awg
- Ageing causes poor sleep and impairs memory go.nature.com/hppcay
- Aphrodisiac craze endangers Himalayan caterpillar go.nature.com/yegirz

SLIDESHOW

NATURAL HISTORY MUSEUM

Letters of Alfred Russel Wallace go online go.nature.com/jepahh

reductions — perhaps even the 17% cut by 2020 that Obama promised at the United Nations climate summit in Copenhagen in 2009.

Many see the reductions as an opportunity. They "should give Americans confidence that climate policies can be effective", says Paul Bledsoe, an environmental consultant in Washington DC and a White House climate-change official under former president Bill Clinton.

As a next step, Obama's administration is expected to impose two greenhouse-gas regulations targeted at power plants, which are responsible for roughly 40% of US emissions. The first, proposed last year by the Environmental Protection Agency but not yet finalized, would limit emissions from new plants, effectively banning the construction of coal-fired plants that are not equipped to capture and sequester carbon dioxide.

A second rule, not yet released, could set emissions limits for existing plants, encouraging the shift towards natural gas. Other rules could target the oil and gas industry by limiting emissions from refineries and drilling sites.

But these piecemeal regulatory efforts will not be sufficient to reduce emissions by 83% by mid-century — a target promised by Obama at the Copenhagen talks. One question is whether the president can build support for a broad programme of energy research and development that could drive down the cost of large-scale, low-carbon energy, and ultimately make a carbon tax or a cap-and-trade agreement politically palatable.

The President's Council of Advisors on Science and Technology has recommended increasing spending on energy research and development from around US$4 billion per year to $16 billion, and some organizations have advocated even more. Armond Cohen, executive director of the Clean Air Task Force in Boston, Massachusetts, argues that Obama could attract conservative support for a strategic research programme focused on large-scale energy technologies such as carbon capture and storage methods and advanced nuclear reactors. Such a programme might look like the energy department's Advanced Research Projects Agency-Energy, itself inspired by a similar defence-department programme, says Cohen. Once technologies are developed, government agencies could use their buying power to expand production and reduce prices.

"We don't want to see Obama walk in and just play small ball again," says Cohen. "Obama really needs to take this innovation problem on head on." ■ SEE EDITORIAL P.577

CORRECTION

The News Feature 'Dynasty' (*Nature* **493,** 286–289; 2013) wrongly stated that Peter Kareiva was a student of Bob Paine. Kareiva is in fact a friend of Paine's.

CLIMATE TEAM CHANGE

Turnover at the top

Even as US President Barack Obama vows action against climate change, he is expected to lose the leaders of three agencies with important stakes in environment issues. The names of possible replacements have begun to circulate, although none has been named officially.

NICOLAS ASFOURI/AFP/GETTY

DEPARTMENT OF ENERGY

Departing: **Steven Chu**

In addition to overseeing US$37 billion awarded to the department by the 2009 US stimulus package, **Chu** (pictured) restructured research at the energy agency, garnering political support for the high-risk, high-reward Advanced Research Projects Agency-Energy, as well as for five Energy Innovation Hubs for integrated and applied research. The stimulus funding came under intense criticism from conservatives, especially the $535 million that went to now-defunct solar-cell manufacturer Solyndra of Fremont, California. But scientists and environmentalists are pushing for an expanded effort to nurture low-carbon technologies.

Candidates: **Byron Dorgan**, **Dan Reicher**

A former Democratic senator for North Dakota, **Dorgan** has three decades of congressional experience representing a state at the heart of the shale-oil boom, and has said that hydraulic-fracturing technologies, used properly, are safe. **Reicher**, an attorney by training, previously headed Google's $1-billion initiative for investing in energy and climate, where he guided investments into solar technologies and electric transport. He served as the energy agency's assistant secretary for efficiency and renewable energy under former president Bill Clinton and was a staff member on then-president Jimmy Carter's commission to investigate the 1979 Three Mile Island nuclear accident in Pennsylvania.

KEVIN WOLF/AP PHOTO

ENVIRONMENTAL PROTECTION AGENCY

Departing: **Lisa Jackson**

On entering office in 2009, **Jackson** (pictured) laid the groundwork for climate regulations by formally declaring carbon dioxide a dangerous pollutant. Since then, her agency has developed the first US greenhouse-gas standards for vehicles, tightened air-quality standards and proposed emissions limits for power plants. Her successor will lead efforts to take action on global warming by imposing new regulations on industry.

Candidates: **Christine Gregoire**, **Bob Perciasepe**

A former governor of Washington, **Gregoire** signed a 2010 law setting up greenhouse-gas reporting requirements and requiring state agencies to reduce emissions, but pulled Washington out of the Western Climate Initiative, a regional emissions-trading programme led by California. She has also been floated as a candidate to lead the Department of the Interior and the energy department. **Perciasepe**, currently deputy administrator at the environment agency, developed a watershed-protection programme while previously at the agency under Bill Clinton. Before joining the Obama administration, he was chief operating officer at the National Audubon Society, a conservation organization in New York.

CHIP SOMODEVILLA/GETTY

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

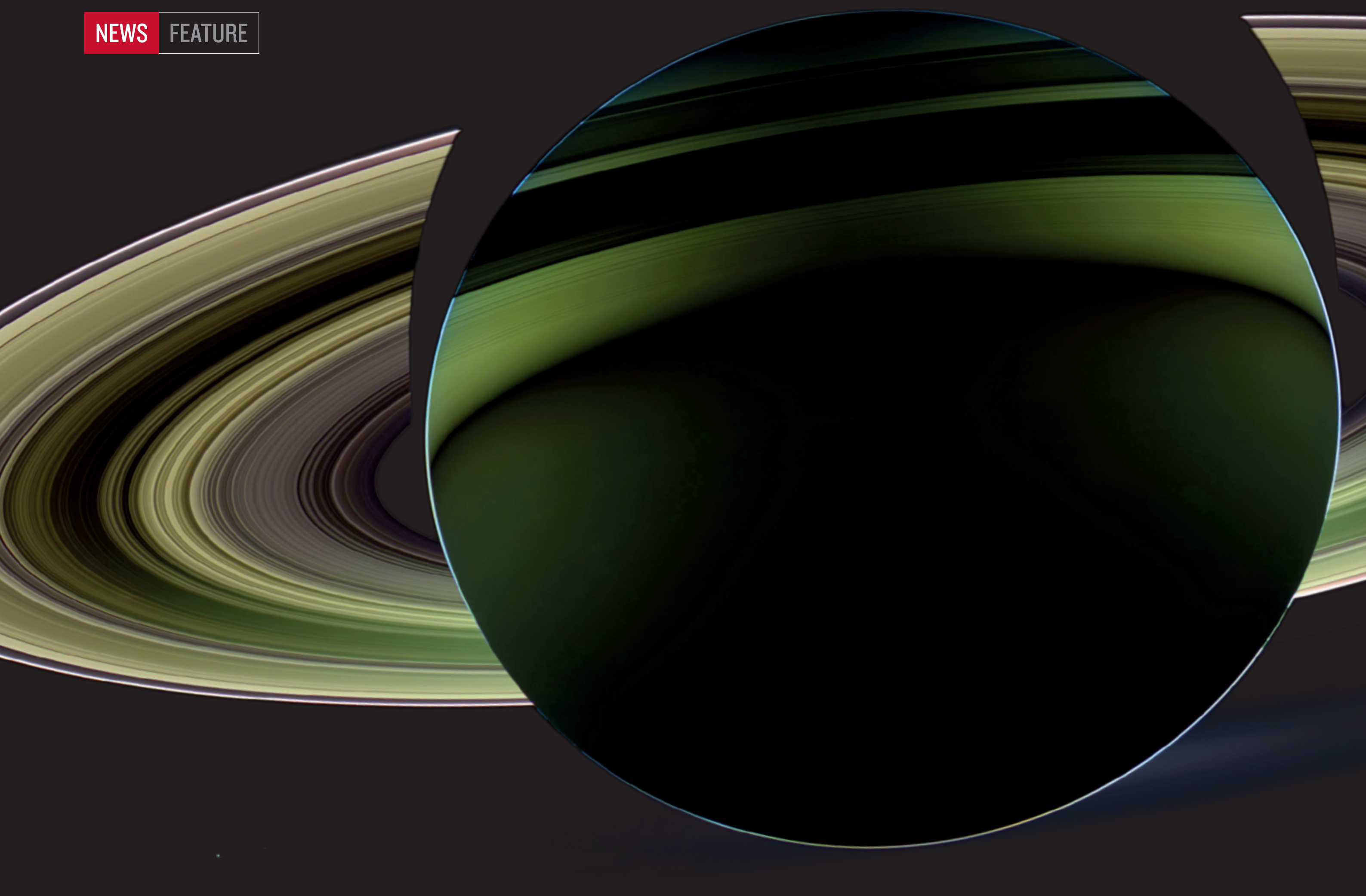
Departing: **Jane Lubchenco**

Lubchenco (pictured) promoted a new US oceans policy and overhauled the way the agency disseminated environmental data. She encountered criticism for the handling of findings related to the Deepwater Horizon oil spill, and she was unable to sell the idea of a federal agency for climate services. Her successor will face questions about catch limits in ocean fisheries, and will need to resolve cost overruns and delays that have plagued weather- and climate-satellite programmes.

Candidate: **Donald Boesch**

A biological oceanographer, **Boesch** is currently president of the Center for Environmental Science at the University of Maryland in Cambridge, where he studies ecosystem management and climate change. He was a member of the White House commission that investigated the 2010 Deepwater Horizon oil spill.

NEWS FEATURE



NASA/JPL-CALTECH/SPACE SCIENCE INST.

CAUGHT IN THE ACT

We may be seeing some of the Solar System's most striking objects during rare moments of glory.

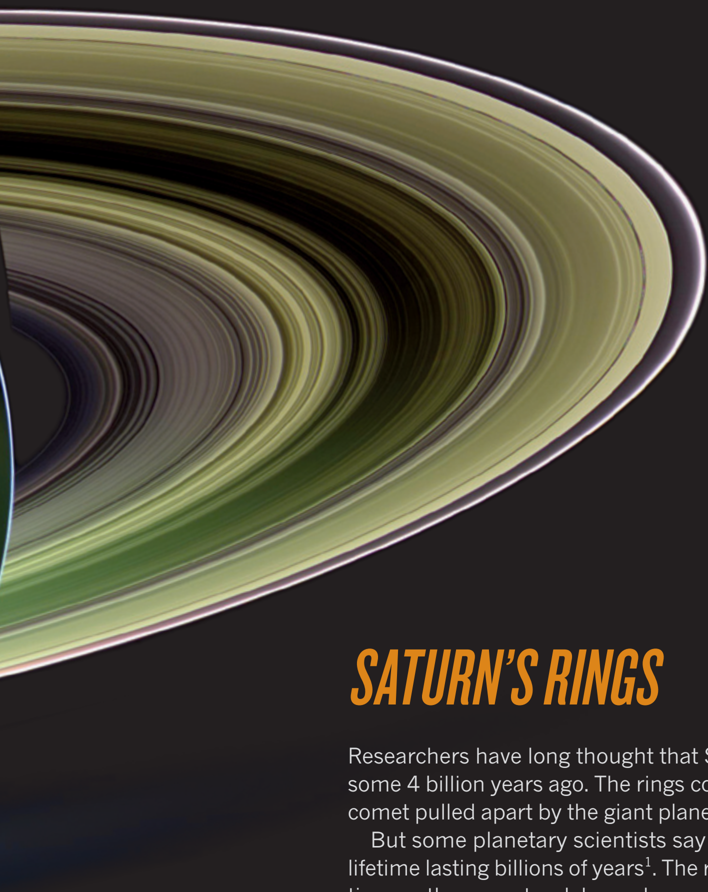
BY MAGGIE MCKEE

Ever since Copernicus evicted Earth from its privileged spot at the centre of the Solar System, researchers have embraced the idea that there is nothing special about our time and place in the Universe. What observers see now, they presume, has been going on for billions of years — and will continue for eons to come.

But observations of the distant reaches of the Solar System made in the past few years are challenging that concept. The most active bodies out there — Jupiter's moon Io and Saturn's moons Enceladus and Titan — may be putting on limited-run shows that humans are lucky to witness. Saturn's brilliant rings, too, might have appeared relatively recently, and could grow dingy over time. Some such proposals make planetary researchers uncomfortable, because it is statistically unlikely that humans would catch any one object engaged in unusual activity — let alone several.

The proposals also go against the grain of one of geology's founding principles: uniformitarianism, which states that planets are shaped by gradual, ongoing processes. "Geologists like things to be the same as they ever were," says Jeff Moore, a planetary scientist at the NASA Ames Research Center in Moffett Field, California. The unchanging world is "philosophically comforting because you don't have to assume you're living in special times", he says.

But on occasion, the available evidence forces researchers out of their comfort zone. Here, *Nature* looks at some of the frozen worlds that may be putting on an unusual spectacle.



NASA/JPL/SPACE SCIENCE INST.

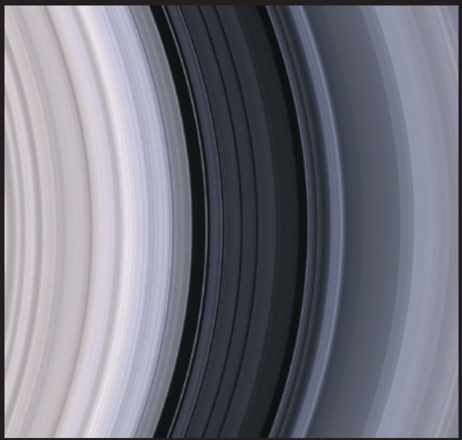
SATURN'S RINGS

Researchers have long thought that Saturn acquired its dazzling adornments early in its life, some 4 billion years ago. The rings could be the glistening remnants of a shattered moon or a comet pulled apart by the giant planet's gravity.

But some planetary scientists say that the rings' resplendence is hard to reconcile with a lifetime lasting billions of years[1]. The rings' particles are 90% water ice and should darken over time as they are struck by carbonaceous dust shed from comets and asteroids. "If you look at the rings of all the other planets — Jupiter, Uranus and Neptune — those rings are all very dark," says Jeff Cuzzi, a planetary scientist at Ames. "That's kind of what you'd expect from heavily polluted material."

According to Cuzzi, the sparkle of Saturn's rings suggests that something — perhaps an icy interloper from beyond Neptune or a large moon of Saturn itself — might have broken apart near the planet and formed the rings within the past few hundred million years, less than 10% of the planet's life so far. The brilliance would be fleeting, because the rings would "get duller and duller" over time, says Cuzzi.

But the idea of young rings presents its own puzzle. Large bodies of the kind that could have formed the bands flew helter-skelter through the Solar System during its first 700 million years or so, but they have grown much rarer since then. There is only a minuscule chance that such a large object whizzed past Saturn in the past one billion years, says Cuzzi. Likewise, he adds, it would be difficult to explain how a moon large enough to form the rings could have fallen close enough to the planet in that time frame.



Saturn's B ring (left) is so bright that some researchers wonder whether it is relatively young.

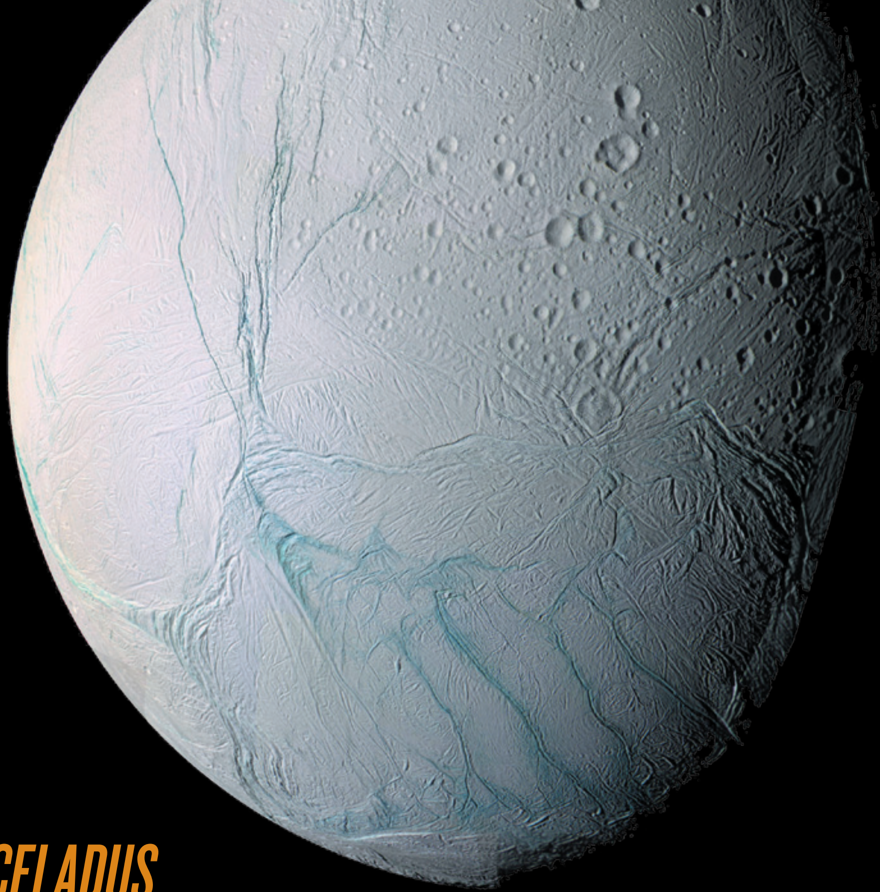
Another possibility is that the rings formed billions of years ago, but somehow retained their youthful glow. That could be the case if they are at least ten times more massive than previously thought, so the dust has so far had little effect. "If you have a thimbleful of black paint, and you drop it into a gallon of white paint, you'll make it pretty dark," says Cuzzi. "But if you drop it into a swimming pool, you won't."

That explanation appeals to Robin Canup, associate vice-president of the planetary-science directorate at the Southwest Research Institute in Boulder, Colorado. "I know of no way to form the rings recently with any reasonable probability," she says.

There is no evidence of any missing mass yet. But it could be hiding in the biggest ring, dubbed the B ring, which is so opaque that researchers cannot study its contents by measuring how light passes through it. The solution to this puzzle could come soon from the Cassini spacecraft, which has been orbiting Saturn since 2004. In 2017, at the end of Cassini's planned lifetime, mission controllers will send it between the planet and the innermost D ring. Comparing the spacecraft's motion at different orbital distances will reveal the rings' mass with unprecedented precision, says Cuzzi.

But Canup warns that "if the Cassini results point to a low mass for the rings, it will be a real mystery".

➔ NATURE.COM
To hear more, listen to the Nature Podcast at:
go.nature.com/mskGod



ENCELADUS

Enceladus's patchwork surface suggests that it has seen bouts of geological activity.

NASA/JPL/SPACE SCIENCE INST.

Enceladus is a fairy moon. As it orbits Saturn, it sprinkles a glittering trail of ice — the E ring — thanks to watery geysers that shoot from its south pole. But researchers have struggled to explain how it can sustain such activity. Enceladus seems to be giving off 16 gigawatts of heat: ten times as much as theorists think it should be able to produce from the decay of radioactive elements in its interior and from the simplest models of tidal heating, the kneading and flexing of the moon caused by Saturn's powerful gravity.

Several explanations have been put forward to account for this furious release of heat, but all rely on arguments that researchers are viewing the moon at a special time. One such proposal, advanced by planetary scientists Craig O'Neill of Macquarie University in Sydney, Australia, and Francis Nimmo of the University of California, Santa Cruz, suggests that over the course of between 100 million and 1 billion years, the internal stresses and strains from tidal forces could build up enough heat to crack the moon's crust, releasing energy and water vapour into space[2].

Jets of water vapour stream from Enceladus's south pole.

NASA/JPL/SPACE SCIENCE INST.

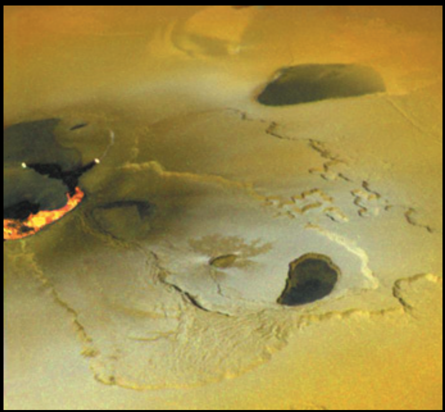
Such activity would last for only about 10 million years before the crust cooled and the geysers died. Then the heat-storage cycle would start anew. "It seems like special pleading — we just happened to catch it in the act," says O'Neill, echoing criticisms that he has heard when presenting the model at conferences. But he points out that the cycle would be just like those of the geysers in Yellowstone National Park in the United States, except on a longer timescale.

Episodic tectonic activity could also explain another discrepancy: why parts of the moon appear to be different ages, with some areas heavily pockmarked by craters and other, fresh-faced regions that have presumably been plastered over by newer crust. A similar patchwork of surfaces is seen on a few other moons, including Jupiter's giant Ganymede and Uranus's small moon Miranda. If these have also gone through cycles of activity, it would make Enceladus less of an outlier. At any given time, there would be a good chance that at least one of them would be passing through a lively period, says O'Neill.

The mystery, then, is why Saturn's moon Mimas, which lies closer to the giant planet than Enceladus and therefore experiences greater tidal forces, shows no sign of tectonic activity. Nimmo says that Mimas may have a different internal composition, making it too rigid to deform, but he acknowledges that this is just one possibility. "Mimas should be producing more heat than Enceladus and it doesn't, and we don't really understand why," he says.

Cassini will collect more clues when it snaps images of Enceladus's south pole between 2015 and 2017, gathering measurements that could refine estimates of the geysers' heat output.

"IT'S POSSIBLE WE SIMPLY DON'T UNDERSTAND THEM."



Io's orbit may be shutting down the moon's volcanoes, such these erupting craters.

ABOVE AND RIGHT: NASA/JPL/UNIV. ARIZONA

IO

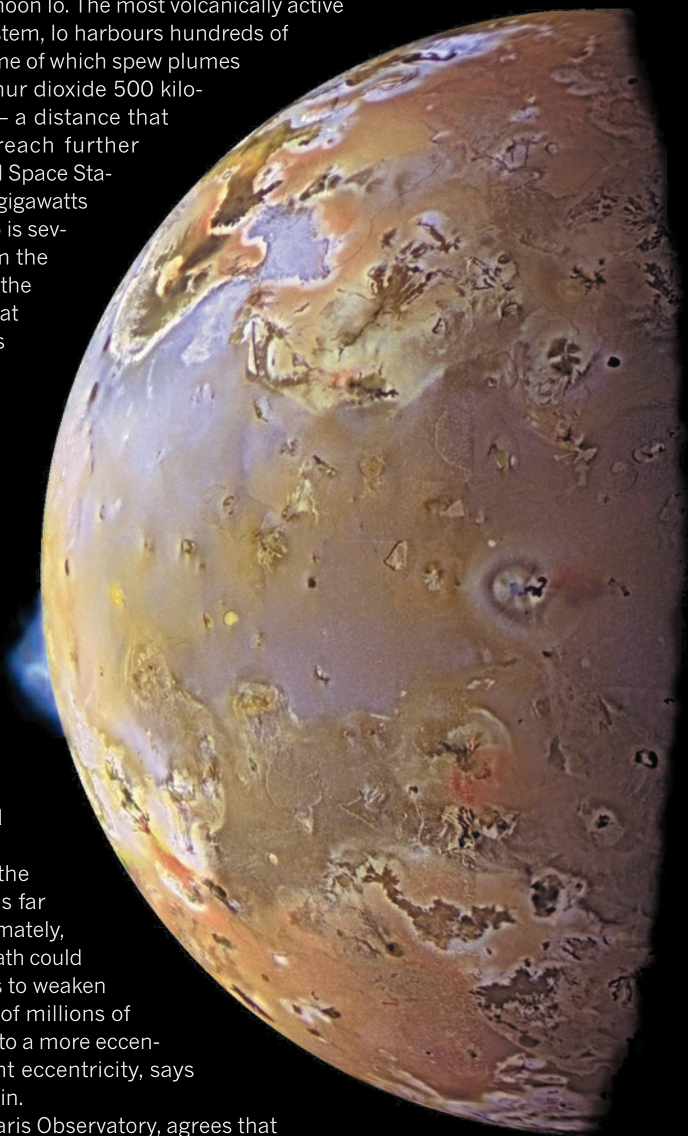
In terms of heat, Enceladus is a firefly in comparison with the furnace of Jupiter's moon Io. The most volcanically active body in the Solar System, Io harbours hundreds of volcanic features, some of which spew plumes of sulphur and sulphur dioxide 500 kilometres into space — a distance that from Earth would reach further than the International Space Station. But the 90,000 gigawatts of heat released by Io is several times more than would be expected from the simplest models of tidal interactions between the moon and Jupiter. That mismatch suggests that "Io is more volcanically active in some periods than others", says David Stevenson, a planetary scientist at the California Institute of Technology in Pasadena.

One possible explanation is that the shape of Io's orbit changes periodically. Io currently takes a slightly elongated, or eccentric, path around Jupiter, thanks to the gravitational influence of two other moons, Europa and Ganymede. Every time Io makes a circuit of Jupiter, the other moons give it a push, "just like a child on a swing", says Stevenson, preventing Jupiter's gravity from pulling Io into a perfectly circular orbit. The eccentric path intensifies the tidal warping, which deforms Io's surface by about 10 metres on each orbit. The frictional heat from all that warping gets released through volcanic eruptions.

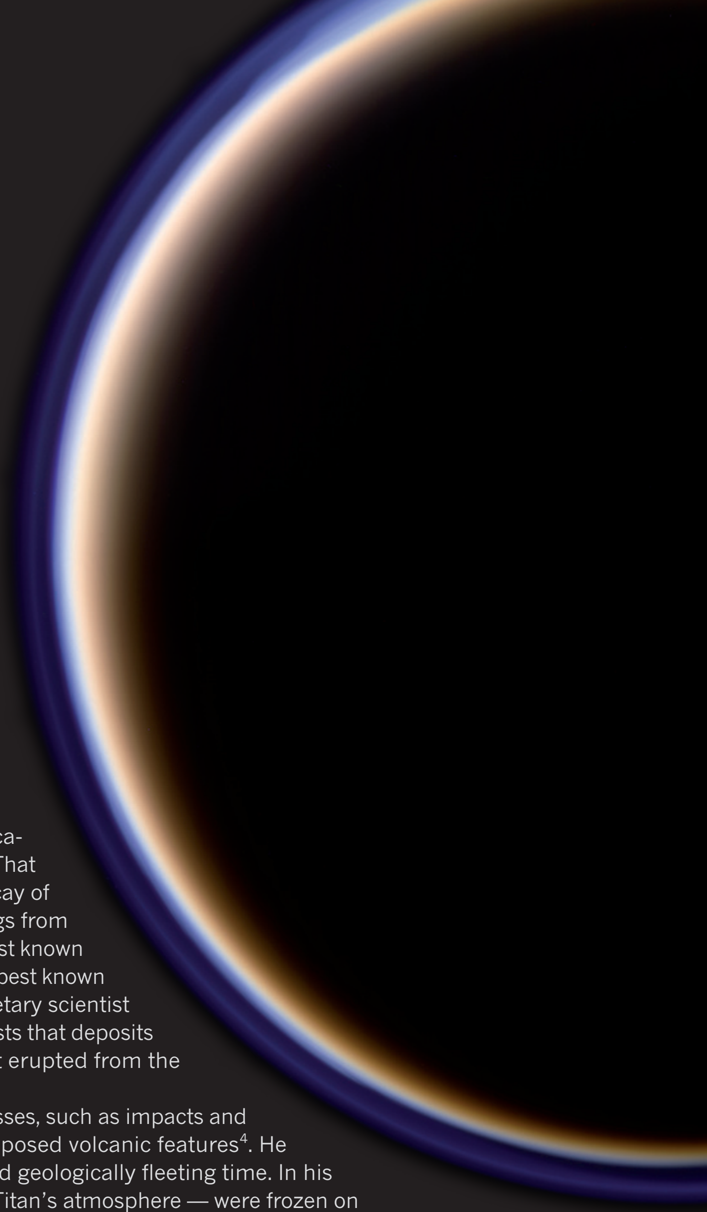
But the same process steals energy from the orbit, so that Io might not be able to swing as far away from Jupiter on subsequent rounds. Ultimately, as energy is drained into internal heating, Io's path could become more circular, causing the tidal forces to weaken and the moon to cool. Then, over the course of millions of years, Europa and Ganymede could push Io into a more eccentric orbit — one with several times its current eccentricity, says Stevenson — and the process could begin again.

Valéry Lainey, a planetary scientist at the Paris Observatory, agrees that there may be cyclical variations in Io's orbit. Some support for that hypothesis comes from observations of Io over more than a century, which show that its orbit may be growing more circular[3]. If so, the moon's raging volcanic activity might be on the wane.

Such orbital transformations "would satisfy the data", says Stevenson. But even though cyclical patterns abound in nature, he says, Io's behaviour, like that of Enceladus, seems so strikingly variable "that it's possible we simply don't understand them".



Io's volcanoes produce sulphurous plumes up to 500 kilometres high.



The hydrocarbon haze in Titan's atmosphere may be a temporary feature.

NASA/JPL

TITAN

When Cassini dropped its Huygens probe through the haze-shrouded atmosphere of Saturn's biggest moon in 2005, it revealed a landscape of sinuous river channels that seems much like Earth's except for one big twist: The liquid that sculpts much of the surface is methane that rains down from hydrocarbon clouds. Yet the atmospheric methane — and its effects on the landscape — ought to be short-lived. Sunlight degrades methane, driving reactions that turn it into heavier hydrocarbons, which should deplete Titan's atmospheric reservoir in a few tens of millions of years. Either researchers are witnessing Titan at a rare moment, not long after a massive release of methane into the atmosphere, or — as many believe — something is replenishing what sunlight destroys.

Cassini revealed a number of what might be ice volcanoes that pump methane up from the moon's interior. That plumbing process could be driven by heat from the decay of radioactive elements inside the moon and from tidal tugs from Saturn. One of these candidate volcanoes is Titan's highest known mountain, Doom Mons, which lies beside the moon's deepest known pit in a region called Sotra Facula. Rosaly Lopes, a planetary scientist at NASA's Jet Propulsion Laboratory in Pasadena, suggests that deposits in that area were formed from methane-rich slush that erupted from the mountain, causing the nearby terrain to collapse.

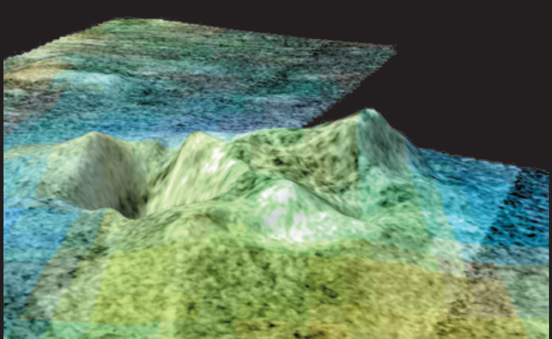
Moore takes a different view, arguing that other processes, such as impacts and erosion by methane rivers, could have created the supposed volcanic features[4]. He thinks that researchers are seeing Titan at a unique and geologically fleeting time. In his view, methane and nitrogen — the main component of Titan's atmosphere — were frozen on the moon's surface until a few tens or hundreds of millions of years ago. At that point, the Sun, which has been growing warmer over its 4.6-billion-year life, vaporized these ices, forming a methane-rich atmosphere within a million years or so.

Methane then condensed from the atmosphere and "rained like hell" over the moon, creating the landscape features, says Moore. Gradually, sunlight turned the methane into heavier hydrocarbons, and the rain tapered off. In another 40 million years or so, says Moore, the methane could completely disappear, and Titan could revert to a nearly unchanging tableau, with blue, nitrogen-filled skies rising above a reddish, hydrocarbon-covered surface.

Ralph Lorenz of the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, argues that Moore's picture is too simplistic. Some evidence suggests, he says, that it would have taken billions of years for the destruction of atmospheric methane to form the hydrocarbon-filled dunes that cover 20% of Titan's surface. If that is so, the liquid-methane cycle has persisted for much of the moon's history.

Continuing observations by Cassini will reveal how much Titan's surface changes on timescales of a few years — allowing researchers to better estimate how long methane rain has been sculpting it.

"I think we have to have a much more nuanced view of Titan through time," says Lorenz. "Titan is bloody complicated." ■



In Titan's Sotra Facula region, the 1.5-kilometre Doom Mons peak (right) sits next to a giant pit (left).

NASA/JPL-CALTECH/ASI/USGS/UNIV. ARIZONA

Maggie McKee *is a freelance writer in Boston, Massachusetts.*

1. Cuzzi, J. N. *et al. Science* **327,** 1470–1475 (2010).
2. O'Neill, C. & Nimmo, F. *Nature Geosci.* **3,** 88–91 (2010).
3. Lainey, V., Arlot, J.-E., Karatekin, Ö. & Van Hoolst, T. *Nature* **459,** 957–959 (2009).
4. Moore, J. M. & Pappalardo, R. T. *Icarus* **212,** 790–806 (2011).

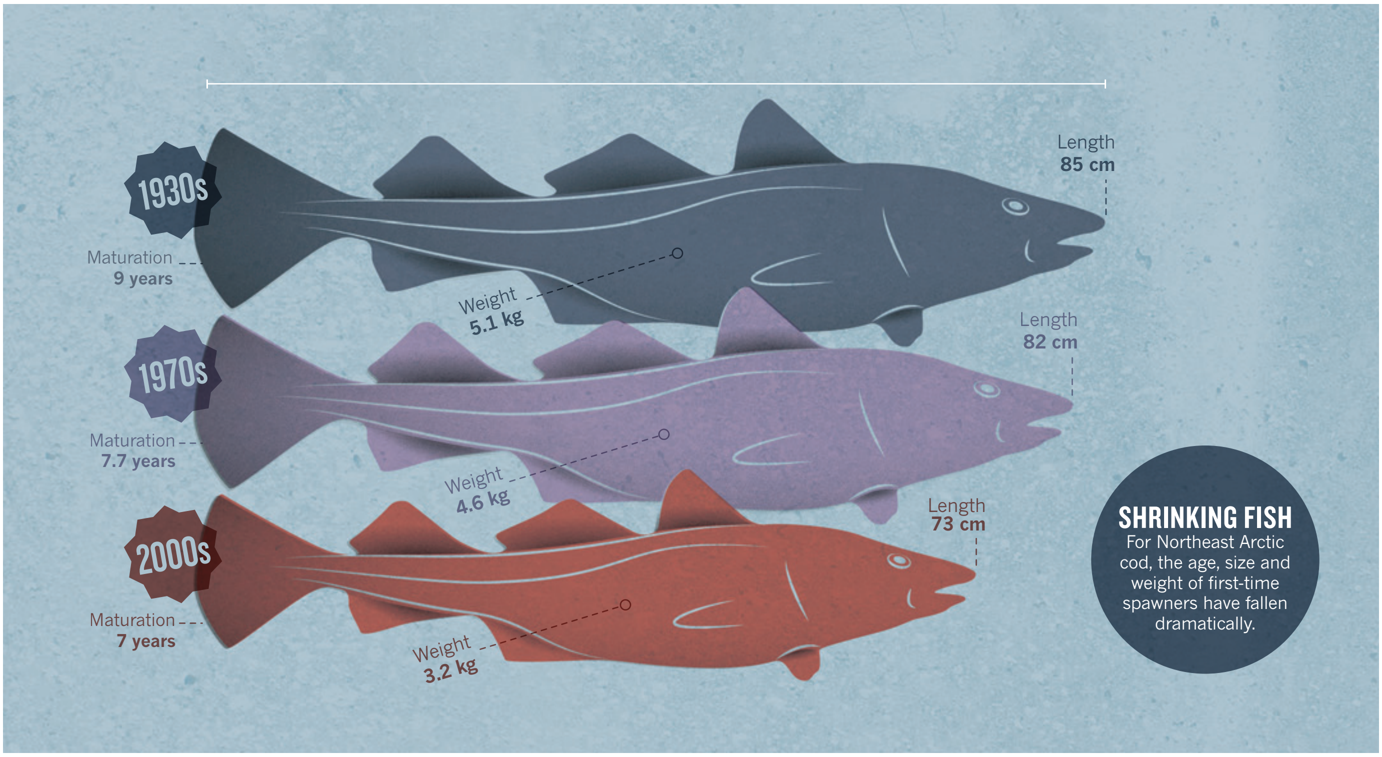


ILLUSTRATION BY WESLEY FERNANDES/*NATURE*; SOURCE: INT. INST. APPL. SYS. ANAL.

A big fight over little fish

Size limits have been a part of fisheries management for decades, but some fear that they are doing more harm than good.

BY BRENDAN BORRELL

One April day, a fisherman named Johan Norman reeled in a female cod near the Norwegian village of Moskenes, where snow-capped mountains rise straight from the sea. He measured the fish: 82 centimetres from the tip of its snout to the tip of its tail. Then he pulled out his knife and sliced off several scales, placing them in a small envelope to deposit at the Institute of Marine Research in Bergen, Norway. The year was 1913.

Over the next century, as those scales sat in a repository, radical changes took place in the world's oceans. The small sailing vessels of Norway and other fishing nations were replaced with industrial bottom trawlers. In 1968, the North Atlantic cod harvest started a precipitous decline, as did other stocks, including salmon, sole and lobster. Then, in the early 1980s, biologists began to report another worrying phenomenon. Fish in some areas were growing more slowly, maturing earlier and laying fewer eggs than before[1]. Not only was this an ominous sign for the sustainability of these fisheries, but smaller fish are less valuable than larger ones because they yield smaller fillets.

Explanations for the shrinking fish have ranged from changes in seawater temperatures to a decline in food resources[2]. But the real culprit could be the practices devised to protect the fisheries. As mandated by various laws and treaties, most trawlers' nets sport a large mesh that allows small, young fish to wriggle free. The reasoning is simple: harvest only the oldest, fattest members of the population and let young fish live to spawn and contribute to the next generation. Fisheries scientists and conservationists support size restrictions because they are thought to protect populations, and fishermen are happy to concentrate on large, high-value fish.

But what if the underlying theory is wrong? Over the past five decades, scientists have come up with little evidence that reducing the catch of juveniles or small fish has improved the annual harvest. Instead, a small chorus of researchers is now arguing, fish are adapting to size restrictions by investing their energy into reaching sexual maturity earlier instead of growing large (see 'Shrinking fish'). And as a result of their small size, they produce fewer eggs. Although these scientists do not deny that overfishing is the greatest threat to fisheries, they say that this evolutionary pressure will have a pernicious impact that will be hard to reverse. "You can safely ignore it for a couple of years, but it's accumulative, so the problem keeps growing," says Mikko Heino, a biologist at the University of Bergen.

The theory is controversial, and many scientists are unconvinced. So last year, Heino turned to Norman's 100-year-old preserved cod scales for help. He extracted DNA from them and is piecing together the whole genome sequence of this fish and others in a hunt for changes in growth and development genes that might explain the species' shrinking size.

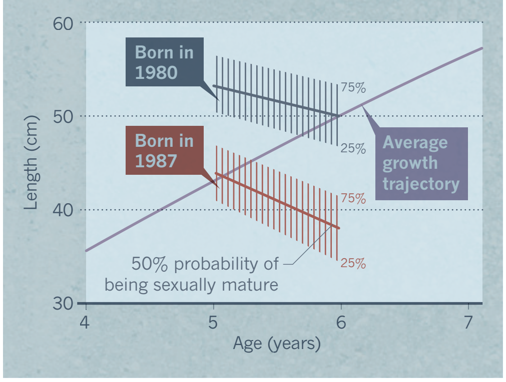
But even if the evolution idea is true, there is some disagreement over what to do about it. Only "a shrinking minority of fools" think that

increasing fishing pressure on juveniles is smart or sustainable, says Carl Walters of the University of British Columbia in Vancouver, Canada.

The theory of fisheries-induced evolution can be traced back to 1981, when the Canadian fisheries scientist William Ricker suggested that coho salmon (*Oncorhynchus kisutch*) and pink salmon (*Oncorhynchus gorbuscha*) were maturing at a smaller size because Japanese gill-net fishermen were targeting only the largest fish on the high seas[1]. By the 1990s, researchers had begun to notice the phenomenon in other species too. But for many years, the consensus was that environmental factors such as climate change and pollution were at play, not genetics.

A SHIFT IN MATURITY

Cohorts of northern cod (*Gadus morhua*) born 7 years apart show that the maturation reaction norm — the probability that a fish will mature at a given size and age — has shifted. Some researchers suspect that evolution is responsible.



SOURCE: REF. 4

Then, in 2002, David Conover and Stephan Munch at the State University of New York in Stony Brook published a contentious experiment[3]. They caught Atlantic silverside (*Menidia menidia*) off the coast of Long Island and established six captive populations of around 1,000 individuals each. After 190 days, they removed 90% of the fish from each population. In the first two populations, they took only the largest fish; in the second two they took only the smallest fish; and in the final two they took individuals of random size. They then stimulated the remaining 10% to breed. After four generations, the fish in the large-harvested populations were about one-third the average weight of those in the random-catch group.

But critics called the experiment unrealistic. The stimulated breeding essentially created a population with a fixed age at sexual maturity, so it was no surprise that removing larger fish favoured those that matured at a smaller size. By contrast, in a natural population, the size at maturity is relatively stable, but age at maturity varies. Slower-growing fish mature later, and faster-growing fish mature earlier. Thus, size limits could select for faster growth, a possibility that Conover and Munch's experiment did not allow. "I was outraged," recalls Walters. "They did an experiment that could only give one result."

PRECOCIOUS COD

The dispute intrigued Heino, a theoretical biologist, who had begun working on his own approach to studying the life history of fish. In the past, researchers would chart a population's maturation reaction norm — the size and age at which fish typically become sexually mature. But Heino realized that comparisons of maturation reaction norms between populations could be misleading if they didn't take into account the variation in growth rates caused by food availability, climate or other environmental factors. So Heino developed a probabilistic approach that considers growth-rate variations.

Using this technique, he showed in a 2004 paper in *Nature*[4] that northern cod (*Gadus morhua*) born in 1987 were maturing at a younger age and a smaller size than those born in 1980, and these changes preceded a dramatic collapse of the species off the coast of Canada in the late 1980s and early 1990s (see 'A shift in maturity').

"It's the most famous fisheries collapse in recent times," says Heino, "You would expect the potential for rapid evolution." Heavy fishing was the main cause of these changes, Heino says, but size-selective fishing compounded the problem. Critics point out that the trend coincided with colder water, heavy sea-ice cover and other factors[2].

Nevertheless, Heino's technique opened up a new field, called Darwinian fisheries management, and evolutionary biologists were soon trying to measure the impacts of size restrictions on other wild populations. A 2009 study[5] used Heino's method to conclude that, of 37 commercial fish stocks, the majority were maturing earlier and at a smaller size than in the past, and that these effects were strongest in heavily fished populations.

Jeff Hard, a geneticist with the US National Oceanographic and Atmospheric Administration Fisheries Service in Seattle, Washington, says that in 1976 the largest class of female salmon — those greater than 100 centimetres in length — accounted for more than 20% of the fish spawning in one Alaskan river. Today, that number is less than 4%, and the number of eggs that females are producing has declined by 16%. But without genetic data from this and other populations, the findings can always be attributed to environmental changes. "It's almost impossible to prove these things," says Andrew Hendry, an evolutionary ecologist at McGill University in Montreal, Canada.

That is why Heino and others are looking to the DNA from historical samples of cod and other species for help. Filip Volckaert of the Dutch-language Catholic University Leuven in Belgium, for example, is sequencing DNA from otoliths, or ear bones, of yellowfin sole (*Limanda aspera*) from every decade back to the 1950s to identify genetic changes that might be linked to growth.

And Heino is complementing the genetic work with his own brand of lab experiment. Inside a special room at his university, he now has nine populations of guppies, and harvests between one-quarter and one-half of the population on the basis of size. To make the experiment more natural than that of Conover and Munch, he allows the guppies to reproduce freely at any age. And, as in nature, the breeding populations contain a wider range of ages and sizes. He expects the experiment, which he started in 2009, to run until 2014.

But it will take a lot to convince the sceptics. "Fisheries-induced evolution is an interesting side issue, but it's been greatly overblown," says Ray Hilborn, a fisheries scientist at the University of Washington in Seattle. There is no question that fished populations are evolving, he says, but some traits, such as earlier age of maturation, may make some fish populations more productive, not less so. The data suggesting that growth rates are slowing are also not yet convincing, he says. The best way to preserve fish populations is simply to fish less, he says.

Heino agrees, but wants to see other changes in marine policy. For example, he does not think that marine reserves should protect only spawning grounds — a common conservation strategy — because that gives another advantage to early-maturing fish, which return to the spawning grounds to breed sooner than late-maturing fish. Second, he says that it is time to abandon most size limits.

Support is growing for these views. Last year, an international group of fisheries experts published a policy paper in *Science*[6] rejecting size limits for a wide range of reasons, including evolutionary issues. Jeppe Kolding of the University of Bergen studies small-scale fishing in Africa, and has found that areas where fishermen use illegal nets that catch large and small fish alike tend to have food webs that are diverse, intact and resemble unharvested areas, only with lower biomass. When fishing pressure is spread across species and sizes, he argues, fishermen can net more fish, yet the risk of wiping out individual populations is lower. "How can you tell me this is a bad fishing method?" he asks.

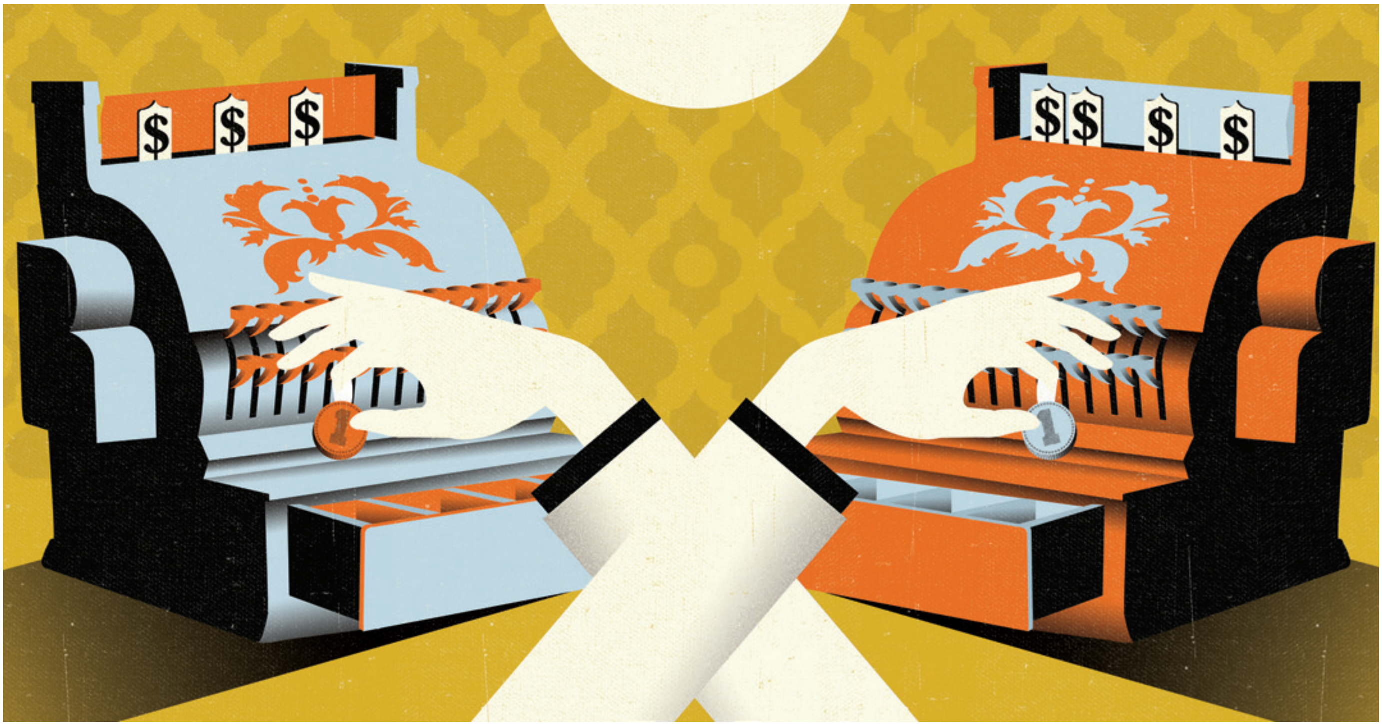
Heino knows that overturning entrenched fishing practices could take decades, and for now he is focusing just on the data. "It requires patience," he says. "The practical implications are something that will keep developing for a long time." ■

Brendan Borrell *is a fellow with the Alicia Patterson Foundation in New York.*

1. Ricker, W. E. *Can. J. Fish. Aquat. Sci.* **38,** 1636–1656 (1981).
2. Kuparinen, A. & Merilä, J. *Trends Ecol. Evol.* **22,** 652–659 (2007).
3. Conover, D. O. & Munch, S. B. *Science* **297,** 94–96 (2002).
4. Olsen, E. M. *et al. Nature* **428,** 932–935 (2004).
5. Sharpe, D. M. T. & Hendry, A. P. *Evol. App.* **2,** 260–275 (2009).
6. Garcia, S. M. *et al. Science* **335,** 1045–1046 (2012).

COMMENT

NEIL WEBB



Same work, twice the money?

Funding agencies may be paying out duplicate grants, according to an analysis by **Harold R. Garner**, **Lauren J. McIver** and **Michael B. Waitzkin**.

With grant success at an all-time low[1], scientists are working harder than ever to fund their research. They respond to the competitive economic times by submitting more applications. They may also simultaneously or serially submit applications to multiple funding agencies to increase their odds of getting funding. Some grant agencies allow the submission of applications with identical or highly similar specific aims, goals, objectives and hypotheses. But we believe that researchers should not accept duplicate funding for the same work — either the whole study or any part of it.

In February last year, the US Government Accountability Office audited the three federal agencies that provide about 94% of all federal funding for medical-sciences research in the United States — the National Institutes of Health (NIH), Department of Defense (DOD) and the Veterans Administration — and found "a potential for unnecessary duplication"[2]. It suggested that the agencies "improve the ability of agency officials to identify possible duplication". The NIH responded to the audit by requiring a detailed evaluation of all proposals from researchers who receive more than US$1.5 million a year in funding to detect any possible "dual/overlapping support"[3]. It has not yet reported any results.

To estimate the extent of double-funding, we systematically compared more than 850,000 funded grant and contract summaries submitted to five of the largest US funders of biomedical research using text similarity software that one author (H.R.G.) invented; we then reviewed a subset of the summaries manually.

We could not determine definitively whether the similar grants we identified were true duplicates — this would require access to the full grant files, which are not available to us without Freedom of Information Act (FOIA) requests. But we did find ▶

evidence that, since 1985 — the earliest year for which grant summaries are available — tens of millions of dollars may have been spent on projects in which at least a portion of the research was already being funded. The problem probably continues today — in the most recent 5 years (2007–11), there were 39 concerningly similar grant pairs, involving more than $20 million. Some of the potential duplicate grants we discovered with our software may have already been identified by the relevant agencies, which may have adjusted the award amount accordingly without updating summaries. But we suspect that there may be many more cases of duplication than our analysis implies.

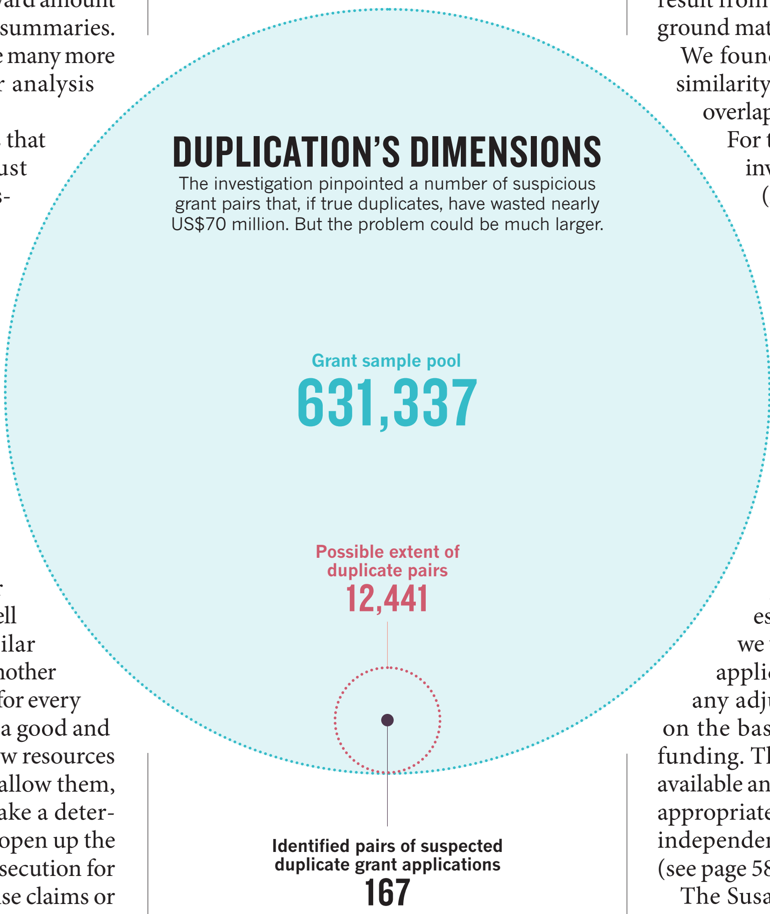
These findings suggest to us that the research community must launch a more thorough investigation into the true extent of duplication. There should be better, clearer and more consistent coordination and guidance about duplication of funding across agencies, both public and private. A central database for all grant proposals would be an excellent first step.

FINDING PAIRS

Government agencies require the disclosure of all current or pending research support, as well as whether the same or a similar proposal is being submitted to another agency. Although explicit rules for every possible scenario do not exist, a good and safe practice is to report any new resources for a study to all funders, and allow them, with this full disclosure, to make a determination. Any violations may open up the grant applicant to criminal prosecution for fraud, civil liability for filing false claims or administrative sanctions, such as debarment from government contracting or suspension as an investigator. Despite these rules, there have been very public cases, often found by serendipity, in which principal investigators have accepted multiple sources of funding for the same project without declaring the existence of other sources[4].

Early last year, we downloaded funded grant summaries (corresponding to 858,717 grants or contracts) from public websites in the NIH, the National Science Foundation (NSF), the DOD, the Department of Energy (DOE) and the Susan G. Komen for the Cure, the largest charitable funder of breast cancer research in the United States. We eliminated more than one-quarter (227,380) of the summaries because they contained fewer than 50 words, so could not be processed accurately by our computational methods. As a result, 2×10^{11} ($(858{,}717 - 227{,}380)^2 / 2$) text-similarity comparisons were possible, so our analysis would only be able to find 54% of all possible duplicates. The number and dates of applications obtained varied greatly across agencies (see 'Double-dip analysis'), with awards totalling more than $200 billion. We recently revisited these databases and found that the DOE had removed the entire database of funded grant summaries.

Our text similarity engine[5], called eTBLAST, calculates a similarity score between each pair of grant summaries using the same approach we established to identify potentially plagiarized scientific literature[6,7]. A two-pass, full-text algorithm accurately and efficiently calculates similarity scores on the basis of the number of shared words placed in the same order in sentences[5]. For this collection of documents, the similarity score for all pairs ranged from 0 to 1.8 (with 1 indicating identical text in two same-length documents, and more than 1 representing identical text in one piece that is longer than the other). We manually reviewed all documents from the federal agencies that received a score of 0.8 or more, and all those from the Susan G. Komen for the Cure that received a score of at least 0.65 — arbitrary cut-offs that in our view provided a sufficiently large sample (1,300 summaries) to capture most of the duplicates. We excluded grants that were obviously associated with an inter-agency combined effort, such as, for example, support for workshops or conferences, large equipment purchases and research involving national laboratory partners.



SAME DIFFERENCE

We focused our manual inspection on comparing specific aims, objectives, goals and hypotheses — because high similarity can result from reuse of introductory or background material.

We found that 11% of the pairs with a similarity of more than 0.8 (or 0.65) had overlapping aims, hypotheses or goals. For these 167 pairs the total money involved was around $200 million (including both grants of the pair) over the entire time records are available. The average size of the first award was 1.9 times that of the potentially overlapping one, so an estimated $69 million of possible overlap funds were found.

Our analysis does not determine whether any likenesses in funded grant pairs are inappropriate, only that the short summaries contained highly similar aims, goals, objectives and hypotheses. To identify true duplicates, we would need to compare the full applications, the awards made and any adjustments made to the awards on the basis of disclosures of duplicate funding. This information is not publicly available and would need to be analysed by appropriate governmental agencies, or by independent groups using FOIA requests (see page 588).

The Susan G. Komen for the Cure did share its private analysis of the highly similar grant pairs we identified (but not the entire applications). At our request, it evaluated 30 pairs, four that it had funded in advance of other agencies, eight that it had funded concurrently and 18 that it had funded subsequent to other agencies. Only four of these pairs had a similarity score of less than 0.8, suggesting that this threshold captured a large fraction (87%) of similar grants.

For two of the pairs, the Susan G. Komen for the Cure reported that its ongoing internal administrative review of grant funding had already identified and made adjustments to funding, which was not reflected in its online summaries. It immediately began to review two active projects with potential overlap that it had not identified. The remaining suspected duplicates would have required the foundation to obtain complete grant applications either from its

archives or from other agencies (using FOIA requests) to make a thorough review, which it did not do.

A better estimate of duplicate funding would account for the grant summaries that were too short to analyse and for those grant pairs with similarity of less than 0.8 that may nonetheless overlap. Using the amount of possible overlap funding ($69 million) inferred from the grants we reviewed, and accounting for the two sensitivities (54% and 87%, the portion of grants captured by the 0.8 threshold), our view is that an exhaustive analysis of all grants could reveal twice as much overlapping funding (69 million/ $(0.87 \times 0.54) = \$147$ million).

Among the summaries with identical or highly similar specific aims, objectives, goals and hypotheses, we found that about 31% ran concurrently. The rest may have been 'recycled', whereby a principal investigator had sent a previously successful grant to another agency. Strangely, the later grant sometimes proposed studies that were cited as preliminary data in the earlier grant, suggesting that the hypothesis had already been resolved and that the proposed research had already long been completed.

With some pairs, a principal investigator seemed to have received a grant that included support for lab members, then sent the grant to another agency for support of graduate students or postdoctoral fellows. There seem to be no clear standards on whether this is an acceptable practice. A full review would be necessary to determine whether the additional funding for the fellow or student on an already funded project was fully disclosed and whether the science project grant budget was adjusted appropriately.

> *"We believe that our analysis may actually have missed duplications."*

We also found similar grants that had different principal investigators (at the same or different institutions) and were funded by different agencies, suggesting that principal investigators may be sharing successful grants with others, thereby enabling them to amplify the amount of funds for a given project, and technically bypassing the formal definition of 'double dipping'. Some of these pairs included current or former co-authors on journal publications, which we discovered from searches of published literature.

In a sampling of around 20 similar grant pairs, we looked in PubMed for publications resulting from the funding, and found that some acknowledged only one agency. Some publications acknowledged additional grant numbers, which, after review, revealed further highly similar grant summaries. These did not meet our 0.8 threshold, but indicate another technique for discovering potential overlap.

DOUBLE-DIP ANALYSIS

After comparing grant and contract summaries with software, all those with a high similarity score were reviewed manually to identify possible duplicates.

Funder	Dates available	Number of applications	Reviewed digitally	Reviewed manually	Suspicious overlaps
Susan J. Komen for the Cure	2003–2011	1,209	1,208	98	30
Department of Defense	1993–2011	10,201	10,086	157	68
Department of Energy*	1995–2009	38,408	9,731	20	3
National Science Foundation	1985–2012	299,332	221,513	446	92
National Institutes of Health	1985–2012	509,567	388,799	579	141
Total		858,717	631,337	1,300	334

*Stopped reporting these data in 2009.

Justifiably, critics will counter that our limited analysis overestimates the problem by not factoring in whether funding agencies have adjusted awards for previous support, and that it suffers from a lack of access to the full grant application and data from all years from all agencies.

However, from our experience in detecting plagiarism using text similarity, we believe that our analysis may actually have missed duplications. For instance, the same comparison techniques have detected plagiarism in 0.04% of biomedical manuscripts[7]. Yet 1.4% of scientists in one survey[8] admitted to plagiarism — that's 35 times the estimated number of duplications in that analysis. If — and it is a very big 'if' — a similar level of duplication did apply in grant applications, the problem could have involved 12,441 pairs of applications (see 'Duplication's dimensions') and up to $5.1 billion since 1985 (or 2.5% of the total funds).

Even if $200 million in duplicated grants represents the full extent of the problem, then some may argue that less than 0.1% of funding since 1985 is too small an amount to warrant concern. But that it is research money that cannot be used to fund the next scientific breakthrough.

GRANT DATABASE

Our findings indicate a need for clearer and more consistent guidance and coordination of grant and contract funding across agencies, both public and private. They may also indicate a need to clarify the standards on what constitutes duplicate funding and to strengthen the surveillance of proposals and funded projects for overlap to ensure adherence to regulations and intent.

We feel that funding agencies and recipient institute administrations could curb duplicate funding more than they are currently doing by using text-similarity comparisons to identify applications and funded grant summaries that warrant closer human scrutiny. That said, similarity software must be continuously updated to respond to changes in grant formats and attempts to evade detection.

Most importantly, creating a central database of grant information from all agencies would enable thorough direct comparisons of all awarded funding, and the prospective identification of similar grant proposals. Such a database, including detailed information within grant applications, and its analysis should remain confidential to ensure that only appropriate facts are released beyond government agencies and their staff. Although this will add some administrative burden, it would help agencies prioritize their awards and target the available dollars more efficiently. ■

Harold R. Garner *and* **Lauren J. McIver** *are at the Virginia Bioinformatics Institute, Virginia Tech, Washington Street, Blacksburg, Virginia.* **Michael B. Waitzkin** *is at Genomeon, Floyd, Virginia.*
e-mails: garner@vbi.vt.edu; mbwaitzkin@gmail.com

1. Kaiser, J. *Science Insider* (20 January 2012).
2. US Government Accountability Office *Report to Congressional Addressees, 2012 Annual Report: Opportunities to Reduce Duplication, Overlap and Fragmentation, Achieve Savings, and Enhance Revenue.* (GAO, 2012); available at go.nature.com/bufrae.
3. Rockey, S. Piloting the $1.5M Special Review. National Institutes of Health Office of Extramural Research Extramural Nexus (8 May 2012). available at go.nature.com/ymffdj
4. Samuel Reich, E. *Nature* **482,** 146 (2012).
5. Lewis, J., Ossowski, S., Hicks, J., Errami, M. & Garner, H. R. *Bioinformatics* **22,** 2298–2304 (2006).
6. Errami, M. *et al. Bioinformatics* **24,** 243–249 (2008).
7. Errami, M. & Garner, H. R. *Nature* **451,** 397–399 (2008).
8. Martinson, B. C. Anderson, M. S. & de Vries, R. *Nature* **435,** 737–738 (2005).

H.R.G. declares competing financial interests. For full details see go.nature.com/q6fkrt.

Editor's note *Nature* is not publishing the grant summaries analysed in this Comment, nor the names of their authors (*Nature*, like most journals, requires Comment authors to make these data available on request). This is because a definitive demonstration of duplication would require access to documents that are not available to the Comment authors. Moreover, the grant authors were not approached for their responses to the analysis. This Comment, although less formal than a *Nature* Letter or Article, underwent peer review.

Scientific genius is extinct

Dean Keith Simonton fears that surprising originality in the natural sciences is a thing of the past, as vast teams finesse knowledge rather than create disciplines.

Many scientists devote their careers to studying phenomena that they can assume will not go away anytime soon. Life forms will always undergo change across generations, so evolutionary biologists will always have a job. But the very phenomenon that I investigate might have actually ceased to exist.

I have devoted more than three decades to studying scientific genius, the highest level of scientific creativity[1]. The creative scientist contributes ideas that are original and useful. The scientific genius, however, offers ideas that are original, useful and surprising. Such momentous leaps — be they theories, discoveries or inventions — are not just extensions of already-established, domain-specific expertise: the scientific genius conceives of a novel expertise.

Albert Einstein's special theory of relativity met these three criteria and required introductory-level textbooks to be rewritten. Einstein overthrew the Newtonian concept of absolute space and time, and revealed a groundbreaking relationship between matter and energy, denoted in his famous equation, $E = mc^2$.

Geniuses have played a decisive part in science in two main ways. First, they have founded new scientific disciplines, such as Galileo's creation of telescopic astronomy. Second, geniuses have revolutionized established disciplines. Charles Darwin, for instance, proposed that species evolve by natural selection at a time when many biologists believed that life forms were fixed from the moment of Biblical creation.

Yet, in my view, neither discipline creation nor revolution is available to contemporary scientists. Our theories and instruments now probe the earliest seconds and farthest reaches of the Universe, and we can investigate the tiniest of life forms and the shortest-lived of subatomic particles. It is difficult to imagine that scientists have overlooked some phenomenon worthy of its own discipline alongside astronomy, physics, chemistry and biology. For more than a century, any new discipline has been a hybrid of one of these, such as astrophysics, biochemistry or astrobiology. Future advances are likely to build on what is already known rather than alter the foundations of knowledge. One of the biggest recent scientific accomplishments is the discovery of the Higgs boson — the existence of which was predicted decades ago.

PETE ELLIS/DRAWGOOD.COM

The days when a doctoral student could be the sole author of four revolutionary papers while working full time as an assistant examiner at a patent office — as Einstein did in 1905 — are probably long gone. Natural sciences have become so big, and the knowledge base so complex and specialized, that much of the cutting-edge work these days tends to emerge from large, well-funded collaborative teams involving many contributors.

SCIENCE OLYMPIANS

At this point, let me add three clarifications. First, I am not saying that scientific progress will cease. On the contrary, I believe that the scientific enterprise will continue to get "faster, higher, stronger". Textbook chapters will continue to be updated. At worst, some disciplines will asymptotically approach some ill-defined limit of precision and comprehensiveness, much as seems to be happening in many competitive sports. Just as athletes can win an Olympic gold medal by beating the world record only by a fraction of a second, scientists can continue to receive Nobel prizes for improving the explanatory breadth of theories or the preciseness of measurements. These laureates still count as 'Olympian scientists'.

Second, I am not arguing that science is becoming 'dumbed down', or that modern investigators are less intelligent than Nicolaus Copernicus, René Descartes, Isaac Newton, Marie Curie or Louis Pasteur. Contemporary scientists generally have very high IQs[2]. If anything, scientists today might require more raw intelligence to become a first-rate researcher than it took to become a genius during the 'heroic age' of the scientific revolution in the sixteenth and seventeenth centuries, given how much information and experience researchers must now acquire to become proficient. It is hard to know whether Pierre-Simon Laplace or James Clerk Maxwell would have been bright enough to master the formidable mathematics of superstring theory, for instance.

Finally, I am not asserting that brilliant scientists can no longer attempt to introduce new paradigms, or even to devise original disciplines. It is just that such innovations seem less likely to catch on. According to Thomas Kuhn's classic analysis of scientific revolutions, a discipline within the physical and biological sciences should not even be receptive to a paradigm shift unless the discipline is in a state of crisis, produced by the accumulation of critical findings that continue to resist explanation[3]. For example, special relativity resolved the impasse set in motion by, among other things, the 1887 experiment by US physicists Albert Michelson and Edward Morley that failed to detect the universal 'ether' assumed to help propagate electromagnetic waves .

Most, if not all, disciplines in the natural sciences do not seem close to this crisis state. The core disciplines have accumulated not so much anomalies as mere loose ends that will be tidied up one way or another. A possible exception is theoretical physics, which is as yet unable to integrate gravity with the other three forces of nature.

Of course, I hope that my thesis is incorrect. I would hate to think that genius in science has become extinct and that my research speciality has become obsolete. It takes only one new scientific genius to prove me wrong. ■

Dean Keith Simonton *is professor of psychology at the University of California at Davis, California 95616, USA.*
e-mail: dksimonton@ucdavis.edu

1. Simonton, D. K. *Scientific Genius: A Psychology of Science* (Cambridge Univ. Press, 1988).
2. Simonton, D. K. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist* (Cambridge Univ. Press, 2004).
3. Kuhn, T. S. *The Structure of Scientific Revolutions* (Univ. Chicago Press, 1996).

CORBIS

A Second World War Marine M-4 tank carries a mine detonator installed by the US Navy Seabees.

MILITARY TECHNOLOGY

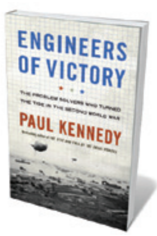
Deadly ingenuity

Two takes on two generations of problem-solving 'geeks of war' fascinate **Sharon Weinberger**.

In November 1940, as the German Luftwaffe was carrying out devastating bombing raids over England, Air Chief Marshal Hugh Dowding wrote, "The war will be won by science thoughtfully applied to operational requirements."

Paul Kennedy offers a slightly different view in *Engineers of Victory*. The noted historian argues that a larger group won: "problem solvers, the scientists and engineers and organizers." A similar crowd stars in Christopher Coker's very different book, *Warrior Geeks*. But whereas Kennedy celebrates such military innovators, Coker — an expert in British defence policy at the London School of Economics — posits that they are depriving us of heroic values of war that date back to ancient Greece.

Kennedy, whose *The Rise and Fall of the Great Powers* (Random House, 1987) covered centuries, narrows his focus in *Engineers of Victory* to a series of complex battles fought across air, land and sea between January 1943 and June 1944. His goal is to identify the neglected "men in the middle", such as Admiral Ben Moreell and Ronnie Harker. Moreell was the founder of the US Navy Seabees, construction battalions whose building sprees included the Tinian airstrip used by the B-29s that delivered the atomic bombs; Harker was a British test pilot whose push to have the American-built P-51 Mustang aircraft outfitted with the British Merlin engine endowed the Allies with aerial superiority.



Engineers of Victory: The Problem Solvers who Turned the Tide in the Second World War
PAUL KENNEDY
Random House: 2013. 464 pp. $30

The book's multilayered descriptions provide keen insight into the complex management that enabled the Allies to win the war. D-Day is a case in point. Kennedy describes the crucial part management played in that victory — by, for instance, coordinating logistics, protecting front-line troops and even ensuring underwater demolition teams could dispose of barbed wire. "Without the middle personnel and the systems they managed, victory would remain out of grasp," writes Kennedy.

Warrior Geeks: How 21st Century Technology is Changing the Way We Fight and Think About War
CHRISTOPHER COKER
C. Hrust and Co.: 2013. 384 pp. £25

Kennedy's love of middle management is somewhat perplexing, however: the scientists in particular often fail to surface in the book. Kennedy is more at ease describing the field of battle than technological innovation. The contribution of the Radiation Laboratory at the Massachusetts Institute of Technology in Cambridge, for example, is mentioned only in passing. Yet Kennedy relishes battlefield arcana, giving the precise number of battleships, cruisers and destroyers for specific engagements, such as the convoy battles of March 1943. This is periodically useful, but at other times evokes US President Barack Obama's stinging criticism of his rival Mitt Romney in a televised debate: "The question is not a game of Battleship, where we're counting ships. It's what are our capabilities?"

What is missing is an analysis of the mechanism behind this middle-management-engineered victory. Kennedy repeatedly throws in the term "feedback loop" — the ability to make improvements in real-time on the basis of new information — as a deciding factor. He argues that the United States and United Kingdom, unlike Japan and Germany, had this kind of flexibility, allowing them to learn from their mistakes. Of course, if this were true across the board, one has to wonder how the Soviets prevailed under Joseph Stalin, whose feedback loop for middle managers often consisted of a bullet to the head.

Yet there is something to Kennedy's argument about adaptability in warfare. Insurgents in Iraq and Afghanistan have proved very adaptable: when coalition forces employed jammers to block the mobile-phone signals used to detonate roadside bombs, they quickly switched to pressure plates and hard wires. By contrast, it took the Pentagon many months to acknowledge that the vulnerable, thin-skinned Humvee vehicles used in Iraq needed replacing.

In *Warrior Geeks*, Coker turns the idea of management on its head. In this fascinating historical and philosophical tour of modern warfare, Coker seizes on some concepts Kennedy mentions only in passing,

such as the introduction during the Second World War of management science and operational research, which went beyond improving weaponry. "The actual use of those weapons and the organization of men using them were seen as scientific problems in themselves," Coker writes of this change. He sees that application, however, as depriving soldiers of their humanity, arguing that the feedback loops lauded by Kennedy are "post-heroic".

The chief concern outlined by Coker is that the ingenuity driving military science is spiralling out of control. The 'geeks' are creating technologies — designer drugs, robotics and neural devices — that, ultimately, he feels, will dehumanize us.

Coker drives home his points with much reference to philosophy and literature, segueing smoothly from trashy Hollywood films such as the forgettable *Stealth* (Rob Cohen, 2005) — rogue drones, anyone? — to the work of the Polish poet Zbigniew Herbert. Sometimes, the philosophizing goes over the top. For instance, Coker sees efforts to develop pharmaceutical interventions to treat post-traumatic stress disorder (PTSD) as scientists wanting to eliminate guilt through drugs. People affected by severe PTSD might argue that such research is in fact about treating symptoms so debilitating that sufferers are often left without jobs or family.

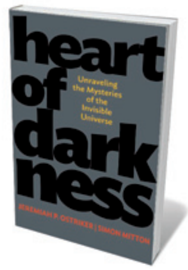
There is much to be said about the dangers of technologically driven warfare, such as the use of armed drones for targeted killings. But the senseless slaughter, in 1994, of more than 500,000 people in Rwanda was carried out in large part by men with machetes. Coker might argue that this form of genocidal warfare was never imbued with Greek values in the first place. But the sheer brutality of that war leaves me doubting that killing someone with the crudest of weapons is any more human, or heroic, than killing by gun-toting robots.

The power of both these books lies in how they prompt us to look through the authors' prisms at the now more than 10-year-old war in Afghanistan. Would empowering Kennedy's problem solvers allow the United States to prevail? Probably not: the building of a modern nation defies managerial or technical solutions.

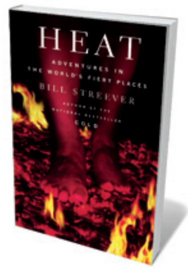
On the bright side, I remain unconvinced that Coker's geeks are going to strip us of our humanity. If that happens, we should blame neither the scientists nor the middlemen, but the politicians who take us into misguided wars in the first place. ■

Sharon Weinberger *is a freelance reporter in Washington DC. She is currently working on a book about the Defense Advanced Research Projects Agency.*
e-mail: sharonweinberger@gmail.com

Books in brief

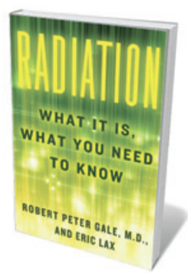


Heart of Darkness: Unraveling the Mysteries of the Invisible Universe
Jeremiah P. Ostriker and Simon Mitton PRINCETON UNIVERSITY PRESS *288 pp. $27.95 (2013)*
In this sweeping chronicle of cosmology, astrophysicist Jeremiah Ostriker and science historian Simon Mitton seamlessly blend historical narrative with lucid scientific explication, from the deeps of classical time to the data-fuelled hyperdrive of the past 50 years. The authors shine what light there is on dark matter and dark energy — a combination Ostriker has helped to pioneer in his models — but admit that the picture is incomplete and plenty of discovery awaits.

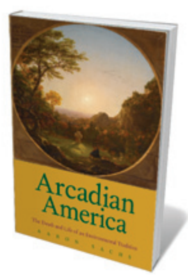


Heat: Adventures in the World's Fiery Places
Bill Streever LITTLE, BROWN *368 pp. $26.99 (2013)*
Biologist Bill Streever might be forgiven for switching from *Cold*, his debut best-seller, to *Heat*: he lives in Alaska. This intense, pacy ride through the thermal kicks off with thirst and ends with quarks freed by heat at the Relativistic Heavy Ion Collider at Brookhaven National Laboratory in Upton, New York. In between, Streever treats us to California wildfires and the chaparral they feed on, John Tyndall's discovery of greenhouse gases, the culinary chemistry of Hervé This, arson, Hawaiian lava fields, atomic bombs, charcoal-burning and even fire-walking. Simmering with verve throughout.

The King of Infinite Space: Euclid and His Elements
David Berlinski BASIC BOOKS *192 pp. $24 (2013)*
Fifty years might be a triumphant span for today's textbooks. Euclid's *Elements* is still fresh after 2,300. As mathematician David Berlinski writes in this pared and elegant homage to the peerless geometer and his magnum opus, the influence of Euclid's axiomatic system remains vast. Berlinski unpacks the axioms, propositions and proofs along with their passage through history — from their influence on Copernicus and Bertrand Russell (who called his encounter with *Elements* "as dazzling as first love") to the non-Euclidean world that sprang open in the nineteenth century.



Radiation: What It Is, What You Need to Know
Robert Peter Gale and Eric Lax KNOPF *288 pp. $26.95 (2013)*
Medical veteran of hot zones from Chernobyl to Fukushima, Robert Peter Gale — haematologist, oncologist and expert in bone-marrow transplants — delivers a guide for those perplexed by radiation. With science writer Eric Lax, Gale weighs up the risks and benefits of industrial, medical and natural radiation clearly, logically and with ample science. But it is Gale's phenomenal frontline experience that gives this book edge — not least a bizarre incident in Goiânia, Brazil, where caesium-137 scavenged from an abandoned radiation-therapy machine eventually affected more than 100,000 locals.



Arcadian America: The Death and Life of an Environmental Tradition
Aaron Sachs YALE UNIVERSITY PRESS *496 pp. $35 (2013)*
From Yosemite to Yellowstone, the US national parks remain a historical touchstone for national environmentalism — but not the only one, argues Aaron Sachs. In a rich mix of history, cultural critique and memoir, Sachs reveals the cemetery as a half-forgotten nineteenth-century landscape tradition. These micro-Arcadias inspired close observation of nature in increasingly urbanized spaces, as well as contemplation of mortality and the sublime.

Thomas Lawrence's portrait of Humphry Davy, who identified nine chemical elements.

ROYAL SOCIETY

HISTORY OF SCIENCE

Elements of romance

Mark Peplow explores chemistry's golden age — and its brushes with Romanticism — at London's Royal Society.

Romantic Chemistry
THE ROYAL SOCIETY, LONDON
Until 14 June 2013

Talk of English Romanticism often conjures up images of William Wordsworth striding through Cumbrian drizzle or Samuel Taylor Coleridge crafting his hallucinatory poem *Kubla Khan*. But this explosion of artistic creativity, concentrated in the first half of the nineteenth century, coincided with a golden age of scientific discovery. More than two dozen chemical elements were identified, and chemistry became the hottest show in town. "It's really the dawning of chemistry as we understand it," says Keith Moore, the Royal Society's head of archives.

Romantic Chemistry, a small exhibition curated by Moore, profiles the scientific stars of the age. Artefacts of their discoveries are displayed alongside cartoons and portraits. Ingots of palladium some two centuries old jostle for space with letters describing pioneering research in beautiful copperplate.

The show also draws out connections between these luminaries and their artistic counterparts. The revolutionary note struck in the United States and France was still sounding for the likes of both Coleridge and chemist Joseph Priestley. And, in other ways, the 'two cultures' barrier had yet to go up.

Some characterize the Romantic movement as a reaction against the growing rationalization of the world through science, as mathematics and measurement began to unlock the secrets of nature. John Keats' lamentation in his 1819 poem *Lamia* sums up this mood: "Philosophy will clip an Angel's wings/Conquer all mysteries by rule and line".

Not so, argues Moore. The chemist and inventor Humphry Davy was friends with Coleridge and Wordsworth, and edited the 1800 second edition of their co-authored *Lyrical Ballads*, the spark that lit English Romantic literature. Coleridge returned the favour with some unabashed public relations in his *Essays on the Principles of Method* in 1818, reproduced in the exhibition: "If in SHAKSPEARE we find nature idealized into poetry … so through the meditative observation of a DAVY, a WOOLLASTON, or a HATCHETT we find poetry … substantiated and realized in nature".

William Hyde Wollaston is not as well known today as Davy, but in the early 1800s his discoveries of palladium and rhodium propelled him to the cutting edge of science. And Wollaston was quick to exploit palladium's commercial potential. On show is an advertising leaflet he circulated, extolling the metal's properties in remarkable scientific detail and noting it as sold "only by Mr Forster, at No. 26 Gerrard St, Soho, London" — an early example of chemical monopoly.

Around the same time, Charles Hatchett claimed to have found a new element in a mineral sample from Massachusetts, which he dubbed columbium. That discovery was confirmed only after 60 years, and it took almost another century before its modern name — niobium — found common usage. Hatchett's 'finder's letter' to the Royal Society is displayed, as is a specimen of niobium ore.

So are letters from Hatchett's assistant, the Irish chemist Peter Woulfe, bearing clues that these Romantic chemists were living through their field's awkward but thrilling teenage years. Perfect, hand-drawn diagrams of distillation equipment that would hardly look out of place in a modern textbook sit alongside lists of chemicals written in the arcane symbols of alchemy.

But the undoubted star of the exhibition is Davy — a romantic among Romantics. His lectures drew London's elite to the Royal Institution and, as James Gillray's satirical cartoon of 1802, *New Discoveries in Pneumatics*, shows, the front row was packed with ladies wearing ostentatious hats, craning eagerly towards the handsome scientist. "Davy was the Brian Cox of his day," smiles Moore, referring to the floppy-haired particle physicist and darling of British science broadcasting.

Indeed, a print of Thomas Lawrence's portrait of Davy shows the chemist dressed in a dandyish shirt with a fashionably high collar, his kid-gloved fist placed firmly on a table. The gleam in his eye would have set his fans swooning. This master of elements could not be more different from Joseph Wright of Derby's *The Alchymist* (1771). A depiction of the seventeenth-century isolation of phosphorus — the first modern discovery of an element — it features a bearded alchemist in a darkened, cluttered room. His face is illuminated by the glow of the element, extracted from copious quantities of urine.

Some 150 years after those crude experiments, in 1820, Davy — who had by then identified nine chemical elements — was appointed president of the Royal Society, following an acrimonious election. The rowdy Romantic chemists had taken over the establishment. Chemistry had arrived. ■

Mark Peplow *is a freelance science journalist based in Cambridge, UK.*
e-mail: peplowscience@gmail.com

HISTORY

Creator — or creationist?

Kevin Padian weighs up a life of a great science popularizer who resisted Darwinism.

Louis Agassiz, a protégé of Georges Cuvier and Alexander von Humboldt, left his native Switzerland for a lecture tour in the United States in 1846. He aimed to boost his reputation, observe the country's geography and wildlife, meet American savants and see their collections. He was so successful an orator on natural history that he ended up with an offer to become a professor at Harvard University in Cambridge, Massachusetts; eventually, his Museum of Comparative Zoology there became the first publicly funded building in the state.

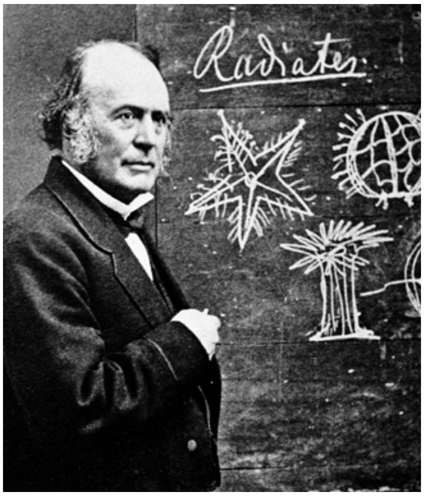
In Christoph Irmscher's balanced and humanistic biography, Agassiz emerges as a genius in natural history and a kind of Svengali for students and the public alike. Yet his stubborn egocentrism eventually undid much of his scientific legacy.

Irmscher, an Americanist who has previously tackled the naturalist John James Audubon and the poet Henry Wadsworth Longfellow, shows an exquisite sensitivity to nature in his portrayal of Agassiz. But his subject, as he acknowledges, does not reveal himself easily. Agassiz spoke incessantly and wrote copiously, but not about his feelings or motivations. So for his insights into the man, Irmscher turned to Agassiz's circle: his wives (the first he treated shabbily and left, the second became his amanuensis), his students and assistants (several of whom eventually aired deep grievances about how he treated them) and his colleague, the botanist Asa Gray (a friend and correspondent of Charles Darwin).

What accounted for Agassiz's popularity? He was a great science communicator, making complex concepts simple and accessible in lectures. He was exotic and authoritative at a time when the United States had no postgraduate scientific institutions. He projected the aura of the great European scholars, and played it to the hilt. And he consoled his audiences by assuring them that life had a purpose, a divine design, discernible to anyone who would undertake to study it assiduously and interpret it according to his teachings. Agassiz excited ordinary people about nature, and they responded with devotion.

He was, however, too grand and overbearing a figure for anyone beneath him to succeed. He was possessive and jealous, and believed that ideas or work produced by anyone he had trained belonged to him. His imperiousness was not unusual, especially for those trained in Europe, where Herr Professor was Lord. Irmscher beautifully details a man at odds with himself, who imagined his actions and motives to be much better than they were. He also shows us — almost reluctantly, because we feel that he wants us to understand his subject as a better man — how Agassiz's views of human races, especially Africans, were even more self-contradictory and poisonous than previous biographers have expressed.

➲ NATURE.COM
For Padian on Darwin, see:
go.nature.com/vppg5w



Louis Agassiz in his prime.

BETTMANN/CORBIS

Why do we still read Agassiz? Well, we don't. By modern standards he was a windbag, well-versed in natural history but full of cock-and-bull about divine guidance of life, and vitriol for those who disagreed. His popular reputation in the mid-Victorian era held even as his scientific reputation declined. That decline can be traced to two factors. Despite Agassiz's prodigious knowledge, he stubbornly rejected evolution and over-insisted on the importance of glaciers in forming geological features. Darwin's view of the world succeeded because he could explain by purely scientific mechanisms, using the facts and literature available to all, the same phenomena that Agassiz (and others, such as the British naturalist Richard Owen) could not.

Here is where I feel Irmscher falls short. He misjudges Darwin as an armchair naturalist, a theorizer, who happily unpacked boxes of specimens that had been collected by others and sent to him at his country sinecure. This is ironic because, although Agassiz was a fine field biologist, he ran a huge campaign to entice the public to send him specimens, which ultimately netted him much more than his museum could hold or organize.

Louis Agassiz: Creator of American Science
CHRISTOPH IRMSCHER
Houghton Mifflin Harcourt: 2013. 448 pp. $35

By contrast, Darwin produced copious works from his five years on *HMS Beagle*; solved long-standing problems in geology, biogeography and natural history long before he returned to England; was elected to several scientific societies within a few years of returning from the *Beagle's* voyage; and was a great natural experimenter and collector whose specimens alone changed forever the ideas of several disciplines. Darwin was able to unpack those boxes, years after he grew too infirm to travel, because he knew precisely whom to ask for them and why they would be important.

The biologist today who doesn't read Agassiz misses some great treatments of glaciology, invertebrates and fishes. The biologist who doesn't read *On The Origin of Species* knows nothing about how evolution works.

More importantly, Irmscher's interest in presenting Agassiz as a sympathetic (not to say justifiable) personality comes at the expense of situating his subject's views in their times and intellectual traditions. We gain only a rough sense of what he thought about evolutionary ideas (and what he proposed in their stead), how his thoughts on embryology and taxonomy were framed philosophically, and how the intellectual traditions he represented (whatever they were, apart from those inspired by von Humboldt) squared with biological thought in Europe and America.

However, philosophy is not the main thrust of this book. Irmscher is a probing and sensitive biographer, the best that Agassiz and his circle could hope for. For a fuller perspective of the man and his times, this should be read with Edward Lurie's *Louis Agassiz: A Life in Science* (Univ. Chicago Press, 1960), Mary Winsor's *Reading the Shape of Nature: Comparative Zoology at the Agassiz Museum* (Univ. Chicago Press, 1991) and Louis Menand's *The Metaphysical Club: A Story of Ideas in America* (Farrar, Straus and Giroux, 2001). ■

Kevin Padian *is professor of integrative biology and curator in the Museum of Paleontology at the University of California, Berkeley.*
kpadian@berkeley.edu

Correspondence

Revived species: how would they survive?

Viewing the revival of extinct species as a laboratory exercise overlooks key behavioural and ecological factors that cannot easily be reproduced (S. Kumar *Nature* **492,** 9; 2012). Hence a recreated dodo might look and feel like one — but it wouldn't quite be a dodo.

Also, re-establishing an extinct species would mean following procedures that are normally used to introduce captive-bred animals to the wild. However, these repopulation attempts have contributed only marginally to biodiversity conservation, largely because the animals do not know how to interact with other members of their species or with their new environment.

Extant species can be trained on the basis of what we have learned from wild individuals, but such information is sparse or non-existent for extinct species. In the absence of their proper ecological niche, 'revived' species reintroduced into the wild would be unlikely to survive.

Diogo Veríssimo *Durrell Institute of Conservation and Ecology, University of Kent, UK.*
dv38@kent.ac.uk
Laure Cugnière *Zoological Society of London, UK.*

Revived species: where will they live?

Subrat Kumar suggests that we should preserve the DNA of vanishing organisms such as tigers so that they can be regenerated later (*Nature* **492,** 9; 2012). But extinctions do not just represent the loss of species — they are the pervasive disintegration and destabilization of ecological networks.

Species are more than the sum of their genes: they are the manifestation of reciprocal interconnectivities between organisms and their environment (C. S. Elton *Animal Ecology* Univ. Chicago Press, 2001). Modern extinctions are an irrefutable symptom of habitat loss and the unravelling of biological processes. If we cannot preserve India's forests and mangrove swamps, for instance, then we cannot save its tigers.

Biotechnology has a role in conservation, but it is not the solution to extinction. Instead, we must protect the integrity of ecosystems and their inherent dynamics. Freezing the tiger's DNA amounts to little more than handing on the responsibility for our actions to the next generation.

J. Grant C. Hopcraft, Markus Borner, Daniel T. Haydon *University of Glasgow, UK, and Frankfurt Zoological Society, Germany.*
grant.hopcraft@glasgow.ac.uk

Concern over US nuclear stewardship

Your discussions of the failure to achieve ignition at the US National Ignition Facility (NIF; *Nature* **491,** 159 and *Nature* **491,** 170; 2012) raise an associated concern about the US Stockpile Stewardship Program that we believe deserves the attention of everyone concerned with the effectiveness of the US nuclear deterrent.

Experimental data from the NIF reveal that lasers can compress hydrogen fuel in fusion capsules, but ignition conditions have not been obtained. Deficiencies in the simulations used to design ignition capsules meant that predictions for when fusion would be achieved were wrong. Worryingly, these deficiencies were not brought to light until experimental data from the NIF made their existence undeniable.

Our concern is that something similar could occur in the Stockpile Stewardship Program, which relies heavily on simulations to assess nuclear-weapons performance. As with ignition, an overly optimistic assessment could result from over-confidence in simulations.

The analogy ends there, because the only experimental data that could definitively expose deficiencies in the nuclear-weapons simulations would have to be obtained from nuclear tests, which are prohibited under a moratorium. The potential consequences would be very serious.

David Sharp, Merri Wood-Schultz *Los Alamos, New Mexico, USA.*
woodschultz@gmail.com

Mauritius threatens its own biodiversity

The unique biodiversity of Mauritius faces a growing threat from an unlikely source: its own government. Last week's meeting of the Intergovernmental Platform on Biodiversity and Ecosystem Services in Bonn, Germany (see go.nature.com/dkyucn), should jolt Mauritius back into honouring its position as the first signatory to the Convention on Biological Diversity.

The Mauritian government is leasing important offshore islet nature reserves for activities that conflict with conservation objectives. This has introduced alien predators and caused the illegal destruction of protected species and habitat, without tangible consequences for those responsible. As a result, two populations of threatened endemic reptiles have already gone extinct from one of the reserves (see go.nature.com/4th4kt).

Under pressure from fruit producers, the government is

also seeking to relax its Wildlife and National Parks Act of 1993 to enable culling of the Mauritian flying fox (*Pteropus niger*), a protected bat species that is classified as endangered by the International Union for Conservation of Nature.

Even some apparently positive actions — such as restoring habitats in the island's national park — are being mismanaged, resulting in rising costs that compromise restoration progress (F. B. V. Florens and C. Baider *Restor. Ecol.* **21,** 1–5; 2013).

F. B. Vincent Florens *University of Mauritius, Réduit, Mauritius.*
vin.florens@uom.ac.mu

Sustain the future by doing more with less

The term sustainability — originally conceived to mean doing more with less — is now used to describe development that meets current needs without compromising those of future generations (World Commission on Environment and Development *Our Common Future*, 1987). As food shortages increase and the global population expands, it is time to revisit the original concept of sustainability.

In accepting the idea of sustainable development as politically correct and all-encompassing, scientists and policy-makers have created a world in which 'sustainability' can be used both to defend and to attack environmental policy. Sustainability needs to be rebranded, for example by shifting consumer focus from greenness to payback and efficiency, and by differentiating between the private costs of policy implementation and the social cost of non-implementation (M. Csutora and Á. Zsóka *J. Consum. Policy* **34,** 67–90; 2011).

The world has changed since 1987, and that 'future generation' has been born. Policies that promote sustainability should aim to provide the best life for as many people as possible — by doing more with less.

Brian G. Fitzgerald *University of Limerick, Ireland.*
brian.g.fitzgerald@ul.ie

Dead language still alive for botanists

I disagree with Frank Udovicic's contention that there is no scientific merit in using Latin, rather than English, for botanical descriptions and diagnoses (*Nature* **492,** 356; 2012). The meaning of descriptive terms in Latin will not change, precisely because it is a dead language. Living languages alter over time and can lead to subtle shifts in interpretation.

For example, the English word 'lavender' can describe either the colour of *Lavandula angustifolia* flowers or a shade of pale purple, whereas the botanical Latin term, *caesius*, has the standardized meaning 'pale blue, with a slight mixture of grey' (W. T. Stearn *Botanical Latin*, 1966).

Both Latin and English diagnoses are permitted under the International Code of Nomenclature for algae, fungi and plants. The International Botanical Congress has refused to make diagnoses in English compulsory. Botanists should therefore be free to use either or both languages.

Adam T. Halamski *Institute of Paleobiology, Warsaw, Poland.*
ath@twarda.pan.pl

Transmission studies resume for avian flu

In January 2012, influenza virus researchers from around the world announced a voluntary pause of 60 days on any research involving highly pathogenic avian influenza H5N1 viruses leading to the generation of viruses that are more transmissible in mammals[1]. We declared a pause to this important research to provide time to explain the public-health benefits of this work, to describe the measures in place to minimize possible risks, and to enable organizations and governments around the world to review their policies (for example, on biosafety, biosecurity, oversight and communication) regarding these experiments.

During the past year, the benefits of this important research have been explained clearly in publications[2–7] and meetings[8–10]. Measures to mitigate the possible risks of the work have been detailed[11–13]. The World Health Organization has released recommendations on laboratory biosafety for those conducting this research[14], and relevant authorities in several countries have reviewed the biosafety, biosecurity and funding conditions under which further research would be conducted on the laboratory-modified H5N1 viruses[10,15–17]. Thus, acknowledging that the aims of the voluntary moratorium have been met in some countries and are close to being met in others, we declare an end to the voluntary moratorium on avian-flu transmission studies.

The controversy surrounding H5N1 virus-transmission research has highlighted the need for a global approach to dealing with dual-use research of concern. Developing comprehensive solutions to resolve all the issues will take time. Meanwhile, H5N1 viruses continue to evolve in nature.

Because H5N1 virus-transmission studies are essential for pandemic preparedness and understanding the adaptation of influenza viruses to mammals, researchers who have approval from their governments and institutions to conduct this research safely, under appropriate biosafety and biosecurity conditions, have a public-health responsibility to resume this important work. Scientists should not restart their work in countries where, as yet, no decision has been reached on the conditions for H5N1 virus transmission research. At this time, this includes the United States and US-funded research conducted in other countries. Scientists should never conduct this type of research without the appropriate facilities, oversight and all necessary approvals.

We consider biosafety level 3 conditions with the considerable enhancements (BSL-3+) as outlined in the referenced publications[11–13] to be appropriate for this type of work, but recognize that some countries may require BSL-4 conditions in accordance with applicable standards (such as Canada).

We fully acknowledge that this research — as with any work on infectious agents — is not without risks. However, because the risk exists in nature that an H5N1 virus capable of transmission in mammals may emerge, the benefits of this work outweigh the risks.

Ron A. M. Fouchier *Erasmus Medical Center, Rotterdam, the Netherlands.*
Adolfo García-Sastre *Icahn School of Medicine at Mount Sinai, New York, USA.*
Yoshihiro Kawaoka* *University of Wisconsin–Madison, Wisconsin, USA, and University of Tokyo, Japan.*
kawaokay@svm.vetmed.wisc.edu
**On behalf of 40 co-authors (see go.nature.com/ed3qkc for a full list).*

1. Fouchier, R. A. M. *et al. Nature* **481,** 443 (2012).
2. Fouchier, R. A. M., Herfst, S. & Osterhaus, A. D. M. E. *Science* **335,** 662–663 (2012).
3. Herfst, S., Osterhaus, A. D. M. E. & Fouchier, R. A. M. *J. Infect. Dis.* **205,** 1628–1631 (2012).
4. Kawaoka, Y. *Nature* **482,** 155 (2012).
5. Yen, H.-L. & Peiris, J. S. M. *Nature* **486,** 332–333 (2012).
6. Morens, D. M., Subbarao, K. & Taubenberger, J. K. *Nature* **486,** 335–340 (2012).
7. Fauci, A. S. & Collins, F. S. *Science* **336,** 1522–1523 (2012).
8. WHO. *Report on Technical Consultation on H5N1 Research Issues* (WHO, 2012); available at go.nature.com/ka2bw4.
9. NSABB. Statement of the National Science Advisory Board for Biosecurity, March 2012; available at go.nature.com/fapzkh.
10. Agenda for Workshop on Gain-of-Function Research on Highly Pathogenic Avian Influenza H5N1 Viruses, 17–18 December 2012, Bethesda, Maryland; available at go.nature.com/tr7r9z.
11. García-Sastre, A. *mBio* **3,** e00049–12 (2012).
12. Imai, M. *et al. Nature* **486,** 420–428 (2012).
13. Herfst, S. *et al. Science* **336,** 1534–1541 (2012).
14. WHO. Guidance on risk control measures for H5N1 transmission research, July 2012; available at go.nature.com/4z4yzg.
15. Public Health Agency of Canada. Advisory on Transmissible H5N1 Viruses, 1 February 2012; available at go.nature.com/jfutoz.
16. COGEM. Letter in response to influenza research at Erasmus MC (in Dutch), 21 March 2012; available at go.nature.com/ef5lix.
17. US Government Policy for Oversight of Life Sciences Dual Use Research of Concern (2012); available at go.nature.com/8rkjap.

Carl Woese

(1928–2012)

Discoverer of life's third domain, the Archaea.

Carl Woese brought a fiercely creative mind, seasoned with rigour, to the biggest questions in biology. By showing almost single-handedly that living organisms fall into three domains — Bacteria, Eukarya and a previously unknown group called the Archaea — he transformed our understanding of how living organisms are related and how they evolved.

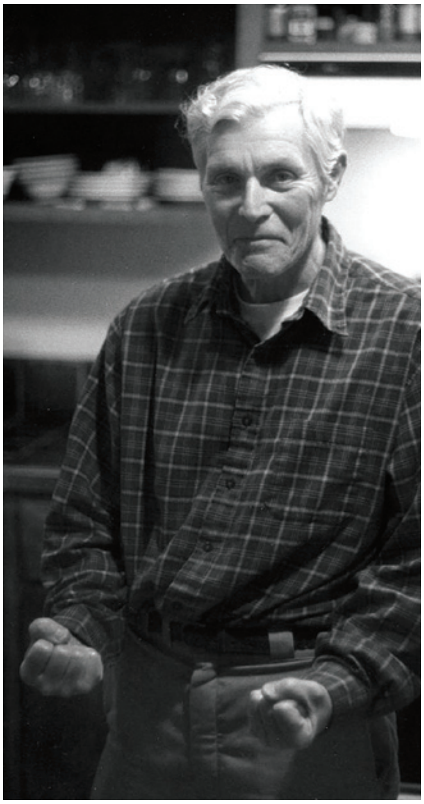
Woese, who died on 30 December, was born in Syracuse in New York in 1928. His undergraduate education was in physics and mathematics at Amherst College in Massachusetts. In 1953, he earned a PhD in biophysics from Yale University in New Haven, Connecticut.

After taking up a research position at the General Electric Research Laboratory in Schenectady, New York, Woese began thinking about the evolution of the genetic code. In 1964, the molecular biologist Sol Spiegelman recruited him to the microbiology department at the University of Illinois in Urbana, where he spent his entire academic career.

At Illinois, Woese examined the nucleotide sequences of 5S ribosomal RNA (a component of ribosomes, which build proteins) from different organisms. He quickly realized that ribosomal RNA is an ideal chronometer for measuring evolutionary distances between living things. It has a slow mutation rate, performs an identical function in all organisms and, because ribosomal RNA interacts specifically with a multitude of proteins, the genes encoding it are unlikely to jump between individuals of different species.

Woese had discovered a window into microbial phylogeny. Until this point, the field had been hopelessly muddy, with identifications of microorganisms based on qualitative characteristics such as differences in shape. In the early 1970s, Woese realized that the sequence of 5S ribosomal RNA contained too few nucleotides (120) to provide a way to classify thousands of organisms. This led him to take on the daunting task of analysing 16S ribosomal RNA, which contains more than 1,500 nucleotides.

Woese began sequencing fragments of 16S ribosomal RNA from every microorganism that he could get his hands on, using RNA 'fingerprinting' — a method developed by British biochemist Fred Sanger. The technique involves separating fragments of RNA in an electric field according to their nucleotide compositions. Woese's enormous undertaking, which involved analysing more than 100 organisms and spanned many years, paid off richly.



One day, the analysis of 16S RNA from a methane-producing organism gave an astonishing result. The familiar pattern of the 100 or so spots, each containing small stretches of RNA, was altered in an unusual way. Several spots present in all bacterial 16S ribosomal RNAs were missing. New spots had appeared, corresponding to ribosomal RNA sequences never seen before. Woese had captured the signature of a different domain of life.

The ribosomal RNAs of some other microorganisms also produced this strange pattern, including those of 'extremophiles', some of which live at temperatures up to 100 °C and secrete sulphuric acid. In 1977, Woese and his postdoc George Fox published their discovery of 'archaebacteria' (now called Archaea) in the *Proceedings of the National Academy of Sciences*, proposing that these organisms were as distantly related to bacteria as bacteria are to eukaryotes.

As well as transforming our understanding of the relationships between living things, Woese's analysis had an impact on ribosome biology. Woese realized that one could use RNA sequences to determine the double-helical folding, or secondary structure, of RNA molecules. Woese and I used this approach to work out the secondary structures of 16S and 23S ribosomal RNA. These comparisons identified the nucleotides in ribosomal RNA that are universally conserved — and therefore crucial to its function — at a time when many believed that the RNA served merely as a structural scaffold for ribosomal proteins.

HARRY NOLLER

Woese's work also spawned a new branch of microbiology: the use of sequence analysis to study natural microbial populations. Combining phylogenetic sequence analysis and the polymerase chain reaction — which amplifies DNA fragments into thousands or millions of copies — makes it possible to identify the microbes in samples from any source, including the ocean and the human body.

At first, Woese's discovery of the Archaea was met with scepticism and even hostility. This, combined with Woese's view of himself as a rebellious outsider, resulted in an often polemical writing style. He took on adversaries as formidable as microbiologist Roger Stanier, taxonomist Ernst Mayr and even Charles Darwin. Yet Woese eventually received the recognition he deserved, including the Crafoord Prize in Biosciences from the Royal Swedish Academy of Sciences in 2003.

Carl once confided to me that a key to his success was "the principle of dynamic incompetence". Visitors to Carl's lab were certainly impressed by his indifference to the mountain of unopened post. His wife Gabriella became so concerned that she persuaded him to let her open the envelopes; among them, she found one with a months-old Dutch postmark. The letter informed Carl that he had been awarded the Leeuwenhoek Medal by the Royal Netherlands Academy of Arts and Sciences — an honour that is given only once a decade and that he shares with Louis Pasteur.

Carl will be deeply missed by colleagues, friends and family. His impact on our understanding of biology is irreversible. ■

Harry Noller *is professor of molecular, cell and developmental biology and director of the Center for Molecular Biology of RNA at the University of California, Santa Cruz, USA.*
e-mail: harry@nuvolari.ucsc.edu

FORUM Genetics

A social rearrangement

Some worker fire ants will tolerate multiple queens in their colony, but others only one. It turns out that this behaviour is governed by a gene cluster on an unusual pair of chromosomes. Two scientists describe what these findings mean to the fields of social evolution, genetics and beyond. SEE LETTER P.664

THE PAPER IN BRIEF

- The two social forms of the fire ant (*Solenopsis invicta*) are under genetic control and follow a Mendelian inheritance pattern that is associated with variants of a single gene, *Gp-9*.
- On page 664 of this issue, Wang *et al.*[1] show that *Gp-9* lies within a single gene cluster containing multiple genes*.
- This 'supergene' is located on a heteromorphic pair of chromosomes, which differ in sequence and structure, in a similar manner to the X and Y sex chromosomes.
- Recombination (the shuffling of DNA between paired chromosomes during cell replication) is suppressed at a region containing more than half of the genes on this chromosome pair.

Genes and queens

ANDREW F. G. BOURKE

One momentous day in the early 1930s, a ship bearing stowaways docked at Mobile, Alabama. On board was a party of fire ants (*Solenopsis invicta*), inadvertently transported from the species' native range in South America. The fire ant has since become a notorious invasive pest, inflicting painful stings, hampering agriculture and damaging native fauna over much of the southern United States and, more recently, reaching Australia and China[2]. It has also provided one of the best case studies of the genetic basis of social behaviour[3,4]. Wang and colleagues' impressive study shows that the fire ant's genome bears its own secret cargo — the first supergene known to be associated with variable social structures within any species of animal.

Like many ants, fire ants live in two social forms (Fig. 1). Monogyne and polygyne colonies contain, respectively, a single queen and multiple queens. In addition, the queens of monogyne colonies are larger and more fecund than polygyne queens. The fire ant's social polymorphism is associated with variation at a single chromosomal location, or locus, containing the gene *Gp-9*, which encodes an odorant-binding protein[3,4]. The two forms of the gene, the alleles *B* and *b*, predict colony type by influencing worker-ant behaviour. Workers with genotype *BB* live under only a single queen, whereas *Bb* workers accept many queens, but only if these queens are *Bb*. This is because, remarkably, *Bb* workers recognize and kill any *BB* queens that they encounter.

The outcome of these behaviours is that the *B* allele occurs in both colony forms but the *b* allele is found in polygyne colonies only. Because *Bb* workers execute all queens lacking the *b* allele, the allele acts as a self-promoting 'green beard' gene (a gene that allows its bearers to discriminate behaviourally between other bearers and non-bearers through an external label)[5]. However, *b* does not spread unchecked, because, in a final twist, it is a lethal recessive allele — in *bb* individuals, the *b* allele causes early death.

Social evolution — the evolution of behaviours that have effects beyond the individual — requires genetic variation to influence social behaviour, so that natural selection has something to act on. The *Gp-9* system shows that, indeed, a multifaceted social trait can be under genetic influence. But how can a single gene have such a wide range of effects? Using next-generation sequencing and other advanced genomic methods, Wang *et al.* confirm previous suspicions[4] that *Gp-9* sits within a supergene that also contains most of the other genes that are differentially expressed between the two colony forms. It is therefore likely that other loci among the more than 600 genes in the supergene, as well as *Gp-9*, contribute to the monogyne–polygyne distinction. But the supergene acts like a single locus because recombination is prevented between the *B* and *b* versions (see 'Chromosome mysteries' below).

Wang *et al.* estimate that the fire ant's supergene arose approximately 390,000 years ago, much more recently than the origin of the fire-ant genus. Other ants show monogyne–polygyne forms, but these need not share a similar genetic basis, because in these species new queens are often admitted to colonies if they are relatives, suggesting that acceptance is not dependent on their similarity at just one locus. The *Gp-9* system therefore seems to be a secondary arrangement that arose after the evolution of polygyny, and long after eusociality (societies with a worker caste). Other complex traits, including some that stem from self-promoting genetic elements[6] (the mouse *t* haplotype is a good example), also involve single, non-recombining, multigenic regions. By using new tools to cast light into the dark hold of the nuclear genome, Wang *et al.* have shown that supergenes can underpin both social behaviour and social structure.

Andrew F. G. Bourke *is in the School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.*
e-mail: a.bourke@uea.ac.uk

Chromosome mysteries

JUDITH E. MANK

At first glance, it might be difficult to see what, if anything, fire-ant behaviour can tell us about the evolution of sex chromosomes. However, Wang and colleagues' study shows us that ant behavioural differences are controlled by a group of genes linked together on 'social chromosomes' that are, in many ways, similar to X and Y chromosomes. Moreover, the fire-ant social chromosomes offer interesting clues about how Y chromosomes initially form.

Y chromosomes are strange, mysterious things. Although initially identical to their X-chromosome partner, they diverge when recombination between the two chromosomes ceases. But because the Y is always paired with

*This article and the paper under discussion[1] were published online on 16 January 2013.

ALEX WILD/VISUALS UNLIMITED, INC./SPL

Figure 1 | Fire ants and queen. Wang and colleagues' genetic analysis[1] of the fire ant *Solenopsis invicta* has revealed a pair of 'social chromosomes' containing a non-recombining region that is expected to encode many of the behavioural traits that define these ants' social structures. The similarity between these chromosomes and sex chromosomes may help our understanding of how Y chromosomes evolve.

the X, halting recombination between them means that the Y stops recombining entirely, although recombination continues between pairs of X chromosomes in females. This causes all sorts of problems for the Y chromosome, such as gene-function decay and the accumulation of repetitive DNA.

However, even though sex chromosomes have been objects of scientific obsession for decades, we do not really understand how recombination between X and Y chromosomes is suppressed. There are theories, of course, the most accepted of which suggests that Y-chromosome inversions, in which a region is flipped end-to-end, are selected for when they encompass both the male sex-determining gene and a nearby gene with male-specific benefits[7]. Such beneficial changes ensure that these genes are transmitted as a single unit — a 'male' supergene — from father to son, as inversions cannot pair correctly during meiosis and therefore recombination is halted in the inverted region between the Y and X. Over time, a series of inversions could theoretically encompass the entire Y chromosome.

There is circumstantial evidence to support this model, namely, in the existence of 'strata' within sex chromosomes that seem to correspond to specific inversion events[8]. However, it has proved exceedingly difficult to identify alleles with sex-specific benefits and, without this, it is almost impossible to find direct evidence for the inversion theory of sex-chromosome evolution.

Enter the fire ants. It was previously recognized that their monogyne and polygyne social forms corresponded to their allele status at the *Gp-9* locus. But these social forms come with an assemblage of morphological and life-history differences, so it is probable that other genes are involved. This begs the question of how alleles at multiple genes can be transmitted as a single unit along with *Gp-9*. Wang and colleagues show that this is accomplished through at least one massive inversion on the chromosome that encompasses the *Gp-9* locus as well as most of the other genes that show expression differences between the social forms. This inversion has, in effect, created a pair of social chromosomes. The inversion prevents recombination between the *B* and *b* forms of the social chromosome, in much the same way as we think inversions might prevent recombination between X and Y chromosomes. It also allows for the transmission of a supergene that encodes the polygyne social structure, directly analogous to the male supergene on the Y chromosome.

The fact that *bb* ants do not survive long enough to reproduce means that the *b* social chromosome is always paired with the *B* chromosome, much like sex chromosomes. And, again just as in the X and Y chromosomes, when recombination is halted between the *B* and *b* chromosomes, the *b* chromosome stops recombining altogether within the inversion. Interestingly, the *b* chromosome exhibits several characteristics also observed on Y chromosomes, including the accumulation of repetitive elements and gene-function decay. So it seems that ant behaviour has a lot to tell us about sex-chromosome evolution after all. ■

Judith E. Mank *is in the Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.*
e-mail: judith.mank@ucl.ac.uk

1. Wang, J. *et al. Nature* **493,** 664–668 (2013).
2. Ascunce, M. S. *et al. Science* **331,** 1066–1068 (2011).
3. Krieger, M. J. B. & Ross, K. G. *Science* **295,** 328–332 (2002).
4. Gotzek, D. & Ross, K. G. *Q. Rev. Biol.* **82,** 201–226 (2007).
5. Keller, L. & Ross, K. G. *Nature* **394,** 573–575 (1998).
6. Burt, A. & Trivers, R. *Genes in Conflict: The Biology of Selfish Genetic Elements* (Harvard Univ. Press, 2006).
7. Charlesworth, D., Charlesworth, B. & Marais, G. *Heredity* **95,** 118–128 (2005).
8. Lahn, B. T. & Page, D. C. *Science* **286,** 964–967 (1999).

SOLAR PHYSICS

The planetary hypothesis revived

The Sun's magnetic activity varies cyclically over a period of about 11 years. An analysis of a new, temporally extended proxy record of this activity hints at a possible planetary influence on the amplitude of the cycle.

PAUL CHARBONNEAU

Right to the end of his life, the Swiss astronomer Rudolf Wolf (1816–93) sought to establish a causal link between the 11-year cycle of the number of dark patches on the Sun (sunspots) and planetary motions. Through his relentless historical detective work, he reconstructed the time series of sunspot number all the way back to the seventeenth century. Taken quite seriously and quantitatively elaborated upon until the end of the nineteenth century, the idea rapidly fell into disfavour following George Ellery Hale's discovery of the magnetic nature of sunspots[1]. Since then, the origin of the sunspot cycle, or solar magnetic cycle, has been sought in the Sun's interior, where the flow of magnetized fluid can lead to self-sustained dynamo action. Rediscovered periodically ever since (pun intended), nowadays the 'planetary hypothesis' for the solar cycle is

ALEX WILD/VISUALS UNLIMITED, INC./SPL

Figure 1 | Fire ants and queen. Wang and colleagues' genetic analysis[1] of the fire ant *Solenopsis invicta* has revealed a pair of 'social chromosomes' containing a non-recombining region that is expected to encode many of the behavioural traits that define these ants' social structures. The similarity between these chromosomes and sex chromosomes may help our understanding of how Y chromosomes evolve.

the X, halting recombination between them means that the Y stops recombining entirely, although recombination continues between pairs of X chromosomes in females. This causes all sorts of problems for the Y chromosome, such as gene-function decay and the accumulation of repetitive DNA.

However, even though sex chromosomes have been objects of scientific obsession for decades, we do not really understand how recombination between X and Y chromosomes is suppressed. There are theories, of course, the most accepted of which suggests that Y-chromosome inversions, in which a region is flipped end-to-end, are selected for when they encompass both the male sex-determining gene and a nearby gene with male-specific benefits[7]. Such beneficial changes ensure that these genes are transmitted as a single unit — a 'male' supergene — from father to son, as inversions cannot pair correctly during meiosis and therefore recombination is halted in the inverted region between the Y and X. Over time, a series of inversions could theoretically encompass the entire Y chromosome.

There is circumstantial evidence to support this model, namely, in the existence of 'strata' within sex chromosomes that seem to correspond to specific inversion events[8]. However, it has proved exceedingly difficult to identify alleles with sex-specific benefits and, without this, it is almost impossible to find direct evidence for the inversion theory of sex-chromosome evolution.

Enter the fire ants. It was previously recognized that their monogyne and polygyne social forms corresponded to their allele status at the *Gp-9* locus. But these social forms come with an assemblage of morphological and life-history differences, so it is probable that other genes are involved. This begs the question of how alleles at multiple genes can be transmitted as a single unit along with *Gp-9*. Wang and colleagues show that this is accomplished through at least one massive inversion on the chromosome that encompasses the *Gp-9* locus as well as most of the other genes that show expression differences between the social forms. This inversion has, in effect, created a pair of social chromosomes. The inversion prevents recombination between the *B* and *b* forms of the social chromosome, in much the same way as we think inversions might prevent recombination between X and Y chromosomes. It also allows for the transmission of a supergene that encodes the polygyne social structure, directly analogous to the male supergene on the Y chromosome.

The fact that *bb* ants do not survive long enough to reproduce means that the *b* social chromosome is always paired with the *B* chromosome, much like sex chromosomes. And, again just as in the X and Y chromosomes, when recombination is halted between the *B* and *b* chromosomes, the *b* chromosome stops recombining altogether within the inversion. Interestingly, the *b* chromosome exhibits several characteristics also observed on Y chromosomes, including the accumulation of repetitive elements and gene-function decay. So it seems that ant behaviour has a lot to tell us about sex-chromosome evolution after all. ■

Judith E. Mank *is in the Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.*
e-mail: judith.mank@ucl.ac.uk

1. Wang, J. *et al. Nature* **493,** 664–668 (2013).
2. Ascunce, M. S. *et al. Science* **331,** 1066–1068 (2011).
3. Krieger, M. J. B. & Ross, K. G. *Science* **295,** 328–332 (2002).
4. Gotzek, D. & Ross, K. G. *Q. Rev. Biol.* **82,** 201–226 (2007).
5. Keller, L. & Ross, K. G. *Nature* **394,** 573–575 (1998).
6. Burt, A. & Trivers, R. *Genes in Conflict: The Biology of Selfish Genetic Elements* (Harvard Univ. Press, 2006).
7. Charlesworth, D., Charlesworth, B. & Marais, G. *Heredity* **95,** 118–128 (2005).
8. Lahn, B. T. & Page, D. C. *Science* **286,** 964–967 (1999).

SOLAR PHYSICS

The planetary hypothesis revived

The Sun's magnetic activity varies cyclically over a period of about 11 years. An analysis of a new, temporally extended proxy record of this activity hints at a possible planetary influence on the amplitude of the cycle.

PAUL CHARBONNEAU

Right to the end of his life, the Swiss astronomer Rudolf Wolf (1816–93) sought to establish a causal link between the 11-year cycle of the number of dark patches on the Sun (sunspots) and planetary motions. Through his relentless historical detective work, he reconstructed the time series of sunspot number all the way back to the seventeenth century. Taken quite seriously and quantitatively elaborated upon until the end of the nineteenth century, the idea rapidly fell into disfavour following George Ellery Hale's discovery of the magnetic nature of sunspots[1]. Since then, the origin of the sunspot cycle, or solar magnetic cycle, has been sought in the Sun's interior, where the flow of magnetized fluid can lead to self-sustained dynamo action. Rediscovered periodically ever since (pun intended), nowadays the 'planetary hypothesis' for the solar cycle is

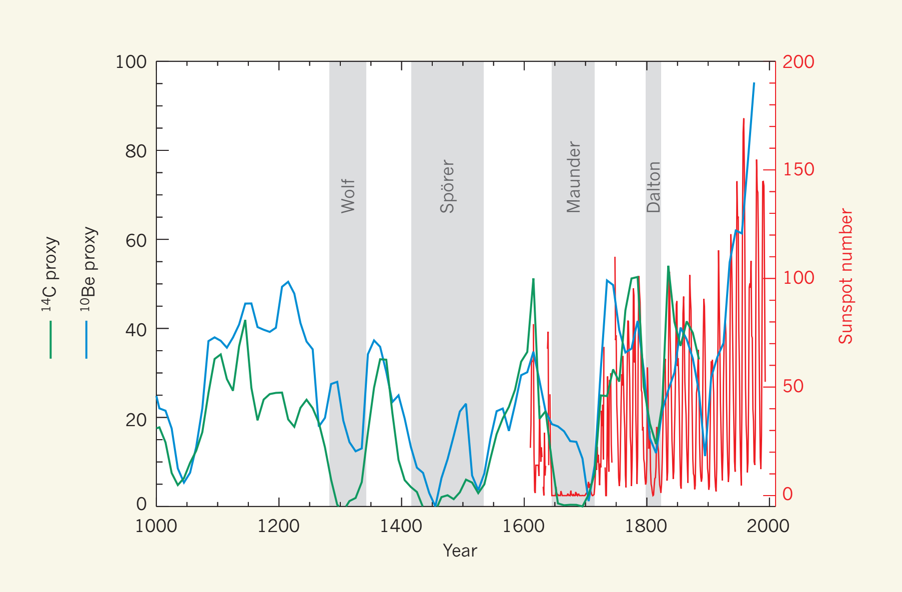


Figure 1 | The sunspot cycle. The graph shows the 11-year cyclical variation in the number of sunspots[11] and two proxy equivalents that are based on the production rate of the radioactive isotopes carbon-14 (^{14}C) and beryllium-10 (^{10}Be). The proxy data, which are averaged over periods of 10 years[12], quite closely follow the secular variation in the sunspot number during the 1600–2000 time interval. The shaded bands show the four intervals of suppressed sunspot activity (known as the Wolf, Spörer, Maunder and Dalton minima) that have occurred in the past millennium. Working with a much longer timespan than that shown here, Abreu *et al.*[2] find that amplitude variations in the sunspot cycle on long timescales exhibit periodicities essentially identical to those that characterize the angular-momentum vector of planetary orbital motions in the Solar System.

usually branded as astrology and summarily dismissed. But writing in *Astronomy & Astrophysics*, Abreu et al.[2] once again revive this hypothesis.

Abreu and colleagues do not question the idea that the magnetic cycle is powered by a hydromagnetic dynamo operating autonomously in the Sun's interior. In fact, they require such an internal dynamo process to provide the basic cycle. But they suggest that its operation is perturbed by the gravitational torque of orbiting planets of the Solar System. In this way, they seek to explain the many well-known centennial-scale quasiperiodicities in solar activity (Fig. 1). Now, generally speaking, such long-timescale modulation can also be produced by conventional dynamo models subjected to stochastic forcing[3], and stochastic perturbations certainly abound in the turbulent solar interior. So why bring planets into the picture at all?

If you propose an unorthodox explanation for a phenomenon already explained by orthodoxy, then the burden of proof is on you. You must do at least as well as orthodoxy at explaining what is already understood, and (it is hoped) succeed in explaining something that the orthodox line cannot. Raising the bar even higher, your novel explanation should be testable in some way within the context of orthodoxy. This adds up to a pretty tall order, but Abreu and colleagues' study meets all three of these requirements.

The authors' data are rock solid and their analysis techniques straightforward and entirely conventional. Working with a 9,400-year-long, high-quality time series for a well-known solar-activity proxy[4], namely, the production rate of the radioactive isotope beryllium-10 (^{10}Be) as determined from ice cores[5], they show that the proxy's time series exhibits many of the same long periodicities as those characterizing the temporal variations of the angular-momentum vector associated with planetary orbital motions. The match is almost perfect for five of the six most prominent periodicities longer than 50 years in the solar-activity record. No purely dynamo-based explanation that I am aware of yields anything remotely close to such an outstanding fit. By using a numerical method known as Monte Carlo simulation, Abreu *et al.* estimate the probability of coincidence for these long periodicities to be less than 1 in 10^6, although this is probably an underestimate, given the properties of the random signals used to test for coincidence between the time series of the ^{10}Be proxy and that of the planetary angular momentum.

Statistical considerations notwithstanding, for such an external torque to have any effect on the solar interior, it must act on an aspherical mass distribution. To explain this, Abreu *et al.* point to the base of the solar convection zone — the Sun's outer shell — where helioseismology studies have detected hints of asphericity[6]. They suggest that the planetary torque acting on the mildly aspherical mass distribution at this depth in the solar interior drives small structural changes that very slightly alter the magnetic-field strength threshold above which buoyant rise sets in and sunspots can form. This could then lead to a large modulation in the rate of sunspot emergence and of the interplanetary magnetic-field strength, in turn modulating the flux of cosmic rays reaching Earth's orbit, and thus the production rate of ^{10}Be.

So, small changes in the Sun's internal structure cause proportionally much larger changes in the amplitude of the solar magnetic cycle — have we now degenerated into astrological homeopathy? Not necessarily. The buoyant destabilization of sunspot-forming magnetic-field concentrations is definitely subjected to thresholds[7], and these are even incorporated in many extant dynamo models of the solar cycle[8–10]. In principle, it should then be a simple matter to carry out dynamo simulations that include a small multi-periodic variation in these thresholds, to assess whether they yield amplitude modulations of the solar magnetic cycle commensurate with those observed in sunspot and proxy data, such as those shown in Figure 1.

This is all pretty far-fetched, but the potential importance of Abreu and colleagues' proposal cannot be overstated. Should it be vindicated, a solid basis for long-term forecasting (and backcasting) of solar activity could then exist. This could greatly benefit current attempts to quantify the past and future long-term influence of solar activity on Earth's space environment, atmosphere and climate.

To sum up, what we have here is a fit to observations unmatched by any other extant explanatory framework, buttressed by a conjectural explanatory physical scenario that is testable at least at some level. It may all turn out to be wrong in the end, but this is definitely not astrology. This is science. ■

Paul Charbonneau *is in the Department of Physics, University of Montreal, Montreal, Quebec H3C3J7, Canada.*
e-mail: paulchar@astro.umontreal.ca

1. Charbonneau, P. *J. Hist. Astron.* **33,** 351–372 (2002).
2. Abreu, J. A., Beer, J., Ferriz-Mas, A., McCracken, K. G. & Steinhilber, F. *Astron. Astrophys.* **548,** A88 (2012).
3. Charbonneau, P. *Living Rev. Solar Phys.* **7,** 3 (2010).
4. Usoskin, I. G. *Living Rev. Solar Phys.* **5,** 3 (2008).
5. Steinhilber, F. *et al. Proc. Natl Acad. Sci. USA* **109,** 5967–5971 (2012).
6. Antia, H. M. & Basu, S. *Astrophys. J.* **735,** L45 (2011).
7. Fan, Y. *Living Rev. Solar Phys.* **6,** 4 (2009).
8. Ossendrijver, M. A. J. H. *Astron. Astrophys.* **359,** 364–372 (2000).
9. Nandy, D. & Choudhuri, A. R. *Astrophys. J.* **551,** 576–585 (2001).
10. Charbonneau, P., St-Jean, C. & Zacharias, P. *Astrophys. J.* **619,** 613–622 (2005).
11. Hoyt, D. V. & Schatten, K. H. *Solar Phys.* **179,** 189–219 (1998).
12. Usoskin, I. G., Solanki, S. K. & Kovaltsov, G. A. *Astron. Astrophys.* **471,** 301–309 (2007).

STRUCTURAL BIOLOGY

Spliceosome's core exposed

The spliceosome complex removes intron sequences from RNA transcripts to form messenger RNA. The structure of a spliceosomal protein, Prp8, reveals the complex's active site and casts light on the origin of splicing. **SEE ARTICLE P.638**

CHARLES C. QUERY & MARIA M. KONARSKA

All eukaryotic organisms share characteristic features, such as the nuclear membrane that separates genetic material from other parts of the cell, and intron sequences that are spliced out from nascent RNA transcripts. Splicing — the excision of introns and the joining of the remaining exon sequences to form a functional messenger RNA — is catalysed by the spliceosome, a multi-component RNA–protein complex composed of subunits called small nuclear ribonucleoprotein particles. On page 638 of this issue, Nagai and colleagues[1] report the long-awaited crystal structure of Prp8, the largest and most evolutionarily conserved protein in the spliceosome[2]. The structure illuminates how the spliceosome might have evolved*.

The presence of large numbers of introns in genes was a driving force in the evolution of complex organisms, because it allowed variation in the way exons were joined together to form mRNA. But the evolutionary origin and the mechanics of splicing have been elusive. Self-splicing RNAs known as group II introns have been postulated to be evolutionarily related to the spliceosome[3,4], because they catalyse similar reactions to those catalysed by the spliceosome and share similar RNA structural elements with it.

Prp8 is a central protein of the spliceosome's catalytic core. It contacts all the substrates and all three of the small nuclear RNAs (snRNAs) thought to form the spliceosome's active site. Despite being highly evolutionarily conserved (61% of the amino-acid sequences of human and yeast Prp8 are the same), its structure has been difficult to obtain. This was partly because, until recently, its protein domains bore little recognizable similarity to any other protein domains for which the structure was known. Furthermore, the massive size of the protein (280 kilodaltons) hampered its analysis.

To obtain crystals for their study, Nagai and collaborators co-expressed two-thirds of the Prp8 protein along with Aar2 — a small chaperone protein that aids the assembly of the U5 small nuclear ribonucleoprotein particle[5], which contains Prp8. The resulting structure of the Aar2–Prp8 complex revealed reverse transcriptase-like (RT) and thumb/X domains, both of which are encoded by many group II introns, and which were recently predicted[6] to be present in Prp8 on the basis of the protein's sequence similarity to other proteins. These were followed by: a linker domain; a domain similar to endonuclease enzymes; a domain similar to RNaseH enzymes; and lastly, a Jab1/MPN domain, typically found in deubiquitinating enzymes. Structures of the latter two domains in Prp8 have previously been visualized using crystallography[7–9].

The structure also reveals that Prp8 forms a loop, with the RT domain at the amino terminus touching the Jab1/MPN domain at the carboxy terminus (see Fig. 2c of the paper[1]). The loop surrounds an internal cavity that almost certainly holds the snRNA components in the spliceosome's catalytic centre. The thumb/X domain forms a platform on which the linker domain folds back, and the RT and endonuclease domains cap the thumb/X–linker on opposite ends (Fig. 1). The linker connects the thumb/X and RT domains, the Jab1/MPN and RT domains, and the thumb/X and endonuclease domains; in large part, it provides a 'lining' on the surface of the internal cavity.

The catalytic activities of reverse transcriptase and of RNaseH depend on amino-acid residues that bind magnesium ions[6–9]. Nagai and colleagues' structure[1], and previous structures of the Prp8 RNaseH domain[7–9], show that the corresponding residues cannot bind magnesium, suggesting that the protein does not have the catalytic activity of those enzymes. Instead, the RT and RNAseH domains of Prp8 are used as structural units, presumably because they provide good surfaces for binding nucleic acids. The perimeter of the spliceosome's active-site core can be delimited by the locations within Prp8 of two crosslinks that form between the protein and the two intron substrates for the first catalytic step of splicing when the spliceosome is irradiated with ultraviolet light: one crosslink

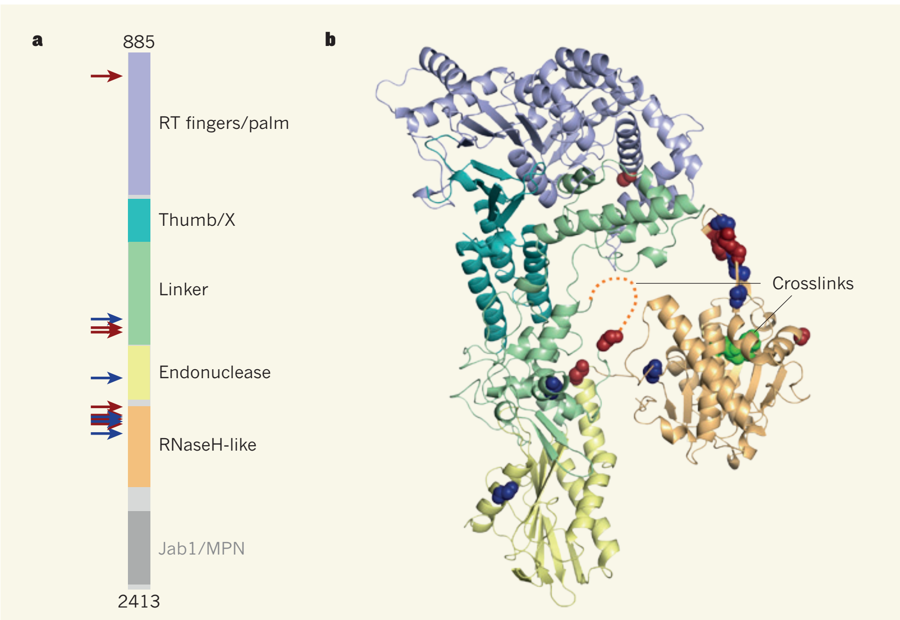


Figure 1 | Model of the Prp8 structure. Prp8 is the largest and most evolutionarily conserved protein of the spliceosome complex, which removes intron sequences from RNA transcripts. Nagai and colleagues[1] report the crystal structure of part of Prp8 (a section consisting of amino acids 885 to 2413) in complex with Aar2, a chaperone protein. **a**, In this representation of the amino-acid sequence of Prp8, the names of domains are shown; colours correspond to those shown in **b**. Blue and red arrows indicate positions of mutations known to affect the first and second steps of splicing, respectively. **b**, In this depiction of Nagai and colleagues' structure, red and blue dots correspond to the mutations shown in **a**. Two crosslinks form between Prp8 and the two substrates for the first catalytic step of splicing when spliceosomes are irradiated with ultraviolet light. The mutations and crosslinks delineate the spliceosome's active-site cavity. The Jab1/MPN domain and the chaperone Aar2 have been omitted for clarity. (Prp8 graphic courtesy of W. P. Galej, Medical Research Council, UK.)

*This article and the paper under discussion[1] were published online on 23 January 2013.

forms in the RNaseH domain[10], and Nagai and colleagues' structure reveals that the other forms in the RT domain. It remains to be seen whether these crosslinks reflect substrate positioning during the first or the second catalytic step of splicing.

In Prp8, many mutations, known as first-step alleles, have been identified that improve the first catalytic step and inhibit the second step, whereas others (second-step alleles) improve the second step and inhibit the first step[11]. Their distribution on the Prp8 structure is remarkable: most first- and second-step-allele mutations line the active-site cavity (Fig. 1), suggesting that they affect interactions with snRNAs and/or with substrates during the catalytic steps. However, some mutations located on the outer Prp8 surface may affect interactions with surrounding components within the spliceosome.

Intriguingly, two strands from the RNaseH domain that form a 'β-hairpin' structure[7–9] carry a large cluster of first- and second-step alleles. This β-hairpin is incorporated into the β-sheet platform that stabilizes contact between the Jab1/MPN and RNaseH domains. The platform is formed by the insertion of a single strand from the Aar2 chaperone into Prp8 (see Fig. 2b of the paper[1]). The requirement of the C terminus of Aar2 to maintain this β-sheet explains the importance of the Aar2 C terminus in the Jab1/MPN–RNaseH domain interaction[5]. This mechanism of β-sheet stabilization is similar to that used by serpin proteins[12], in which a disordered strand from one domain of the serpin undergoes a conformational change to insert into a β-sheet in another domain. In the absence of Aar2, the Jab1/MPN and RNaseH domains might move relative to each other, or else an unstructured region from another protein might help to stabilize the β-sheet.

The authors' work raises even more questions about the highly dynamic spliceosome. How closely does their crystal structure relate to Prp8's structure during catalysis? Aar2 is released during the late stages of the U5 small nuclear ribonucleoprotein particle's biosynthesis, probably when an ATPase enzyme, Brr2, joins U5. Does Brr2 replace Aar2 to maintain a β-sheet-forming interaction with the RNaseH and Jab1/MPN domains? If so, then how much does the structure of Prp8 alter upon this exchange? How are snRNAs and substrates loaded into Prp8's interior cavity, and how does the protein's structure change when the cavity opens for loading and closes for catalysis? And what contacts between domains stabilize the first and second catalytic steps?

Most importantly, Nagai and colleagues' structure stimulates many ideas about the origins of splicing. The domain architecture suggests a close evolutionary relationship between Prp8 and group II introns, which encode maturase proteins that contain RT and thumb/X domains. The authors propose that the RT domain of a maturase encoded by a group II intron co-opted additional domains to create Prp8.

However, the genes of many viruses and retroelements (virus-like RNA sequences that integrate into host genomes) express a single, long 'polyprotein' that is subsequently cleaved into individual functional components. The domain structure of Prp8 might therefore reflect the polyprotein of an ancient retroelement that was the progenitor of group II introns and of the spliceosome. In this scenario, the active sites of the RT and RNaseH domains have been lost, and the catalytic domain responsible for cleaving apart the domains of the progenitor Prp8 have been lost or inactivated and/or were originally contained in the C-terminal Jab1/MPN domain. This composition of functional domains is exactly what would be expected of an ancient retroelement.

Prp8 also contains an endonuclease-like domain. This has retained its catalytic residues, but Nagai and colleagues' mutational analysis reveals that they are not required for splicing and that, notably, the endonuclease active site points away from Prp8's active-site cavity. Might this domain function as a homing endonuclease (an enzyme that enables retroelement genes to move around within host genomes) for the spliceosome, analogous to the retrotransposition behaviour of group II introns? Such activity, in the presence of a reverse transcriptase (the enzyme that generates complementary DNA from an RNA template), could provide a mechanism for inserting introns into DNA. This possibility calls for a reconsideration of the mechanisms by which introns are acquired by genomes. ■

Charles C. Query *is in the Department of Cell Biology, Albert Einstein College of Medicine, Bronx, New York 10461-1975, USA.*
Maria M. Konarska *is at the Laboratory of Molecular Biology and Biochemistry, Rockefeller University, New York, New York 10021, USA.*
e-mails: charles.query@einstein.yu.edu; konarsk@rockefeller.edu

1. Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. *Nature* **493,** 638–643 (2013).
2. Grainger, R. J. & Beggs, J. D. *RNA* **11,** 533–557 (2005).
3. Sharp, P. A. *Cell* **42,** 397–400 (1985).
4. Cech, T. R. *Cell* **44,** 207–210 (1986).
5. Weber, G. *et al. Genes Dev.* **25,** 1601–1612 (2011).
6. Dlakić, M. & Mushegian, A. *RNA* **17,** 799–808 (2011).
7. Pena, V., Rozov, A., Fabrizio, P., Lührmann, R. & Wahl, M. C. *EMBO J.* **27,** 2929–2940 (2008).
8. Ritchie, D. B. *et al. Nature Struct. Mol. Biol.* **15,** 1199–1205 (2008).
9. Yang, K., Zhang, L., Xu, T., Heroux, A. & Zhao, R. *Proc. Natl Acad. Sci. USA* **105,** 13817–13822 (2008).
10. Reyes, J. L., Gustafson, E. H., Luo, H. R., Moore, M. J. & Konarska, M. M. *RNA* **5,** 167–179 (1999).
11. Liu, L., Query, C. C. & Konarska, M. M. *Nature Struct. Mol. Biol.* **14,** 519–526 (2007).
12. Liu, L., Mushero, N., Hedstrom, L. & Gershenson, A. *Biochemistry* **45,** 10865–10872 (2006).

BIOGEOCHEMISTRY

The depths of nitrogen cycling

Breakdown of dissolved organic nitrogen in the ocean had been thought to be the preserve of microbes at the surface. The discovery that these microbes are not up to the task calls for a reassessment of the biogeochemistry of this nitrogen pool.

MAREN VOSS & SUSANNA HIETANEN

The largest pool of fixed nitrogen on Earth is dissolved organic nitrogen in the oceans, but its dynamics and transport between ocean zones are largely unknown. Writing in *Global Biogeochemical Cycles*, Letscher *et al.*[1] report that the observed concentration gradients of dissolved organic nitrogen in the oceans are the result of the interplay between high production in upwelling regions, the transfer of these waters from the surface to lower depths, and degradation by specifically adapted microbial communities. This contradicts the existing theory of the dynamics of nitrogen cycling.

The role of dissolved organic nitrogen (DON) in ocean productivity has long puzzled marine biogeochemists. DON accounts for roughly 60% of the reactive nitrogen in the ocean, which is a much higher percentage than that of readily available nitrogen-containing inorganic nutrients (such as nitrate) and particulate organic matter[2]. DON has therefore been proposed to be one of the major sources of nutrients in the open ocean, and to be responsible for the concentration gradients of nitrate that are observed at different regions and depths[3]. For a long time, DON was regarded as unsuitable for microbial uptake, but we now know that part of it fuels primary production — the formation

forms in the RNaseH domain[10], and Nagai and colleagues' structure reveals that the other forms in the RT domain. It remains to be seen whether these crosslinks reflect substrate positioning during the first or the second catalytic step of splicing.

In Prp8, many mutations, known as first-step alleles, have been identified that improve the first catalytic step and inhibit the second step, whereas others (second-step alleles) improve the second step and inhibit the first step[11]. Their distribution on the Prp8 structure is remarkable: most first- and second-step-allele mutations line the active-site cavity (Fig. 1), suggesting that they affect interactions with snRNAs and/or with substrates during the catalytic steps. However, some mutations located on the outer Prp8 surface may affect interactions with surrounding components within the spliceosome.

Intriguingly, two strands from the RNaseH domain that form a 'β-hairpin' structure[7–9] carry a large cluster of first- and second-step alleles. This β-hairpin is incorporated into the β-sheet platform that stabilizes contact between the Jab1/MPN and RNaseH domains. The platform is formed by the insertion of a single strand from the Aar2 chaperone into Prp8 (see Fig. 2b of the paper[1]). The requirement of the C terminus of Aar2 to maintain this β-sheet explains the importance of the Aar2 C terminus in the Jab1/MPN–RNaseH domain interaction[5]. This mechanism of β-sheet stabilization is similar to that used by serpin proteins[12], in which a disordered strand from one domain of the serpin undergoes a conformational change to insert into a β-sheet in another domain. In the absence of Aar2, the Jab1/MPN and RNaseH domains might move relative to each other, or else an unstructured region from another protein might help to stabilize the β-sheet.

The authors' work raises even more questions about the highly dynamic spliceosome. How closely does their crystal structure relate to Prp8's structure during catalysis? Aar2 is released during the late stages of the U5 small nuclear ribonucleoprotein particle's biosynthesis, probably when an ATPase enzyme, Brr2, joins U5. Does Brr2 replace Aar2 to maintain a β-sheet-forming interaction with the RNaseH and Jab1/MPN domains? If so, then how much does the structure of Prp8 alter upon this exchange? How are snRNAs and substrates loaded into Prp8's interior cavity, and how does the protein's structure change when the cavity opens for loading and closes for catalysis? And what contacts between domains stabilize the first and second catalytic steps?

Most importantly, Nagai and colleagues' structure stimulates many ideas about the origins of splicing. The domain architecture suggests a close evolutionary relationship between Prp8 and group II introns, which encode maturase proteins that contain RT and thumb/X domains. The authors propose that the RT domain of a maturase encoded by a group II intron co-opted additional domains to create Prp8.

However, the genes of many viruses and retroelements (virus-like RNA sequences that integrate into host genomes) express a single, long 'polyprotein' that is subsequently cleaved into individual functional components. The domain structure of Prp8 might therefore reflect the polyprotein of an ancient retroelement that was the progenitor of group II introns and of the spliceosome. In this scenario, the active sites of the RT and RNaseH domains have been lost, and the catalytic domain responsible for cleaving apart the domains of the progenitor Prp8 have been lost or inactivated and/or were originally contained in the C-terminal Jab1/MPN domain. This composition of functional domains is exactly what would be expected of an ancient retroelement.

Prp8 also contains an endonuclease-like domain. This has retained its catalytic residues, but Nagai and colleagues' mutational analysis reveals that they are not required for splicing and that, notably, the endonuclease active site points away from Prp8's active-site cavity. Might this domain function as a homing endonuclease (an enzyme that enables retroelement genes to move around within host genomes) for the spliceosome, analogous to the retrotransposition behaviour of group II introns? Such activity, in the presence of a reverse transcriptase (the enzyme that generates complementary DNA from an RNA template), could provide a mechanism for inserting introns into DNA. This possibility calls for a reconsideration of the mechanisms by which introns are acquired by genomes. ■

Charles C. Query *is in the Department of Cell Biology, Albert Einstein College of Medicine, Bronx, New York 10461-1975, USA.*
Maria M. Konarska *is at the Laboratory of Molecular Biology and Biochemistry, Rockefeller University, New York, New York 10021, USA.*
e-mails: charles.query@einstein.yu.edu; konarsk@rockefeller.edu

1. Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. *Nature* **493,** 638–643 (2013).
2. Grainger, R. J. & Beggs, J. D. *RNA* **11,** 533–557 (2005).
3. Sharp, P. A. *Cell* **42,** 397–400 (1985).
4. Cech, T. R. *Cell* **44,** 207–210 (1986).
5. Weber, G. *et al. Genes Dev.* **25,** 1601–1612 (2011).
6. Dlakić, M. & Mushegian, A. *RNA* **17,** 799–808 (2011).
7. Pena, V., Rozov, A., Fabrizio, P., Lührmann, R. & Wahl, M. C. *EMBO J.* **27,** 2929–2940 (2008).
8. Ritchie, D. B. *et al. Nature Struct. Mol. Biol.* **15,** 1199–1205 (2008).
9. Yang, K., Zhang, L., Xu, T., Heroux, A. & Zhao, R. *Proc. Natl Acad. Sci. USA* **105,** 13817–13822 (2008).
10. Reyes, J. L., Gustafson, E. H., Luo, H. R., Moore, M. J. & Konarska, M. M. *RNA* **5,** 167–179 (1999).
11. Liu, L., Query, C. C. & Konarska, M. M. *Nature Struct. Mol. Biol.* **14,** 519–526 (2007).
12. Liu, L., Mushero, N., Hedstrom, L. & Gershenson, A. *Biochemistry* **45,** 10865–10872 (2006).

BIOGEOCHEMISTRY

The depths of nitrogen cycling

Breakdown of dissolved organic nitrogen in the ocean had been thought to be the preserve of microbes at the surface. The discovery that these microbes are not up to the task calls for a reassessment of the biogeochemistry of this nitrogen pool.

MAREN VOSS & SUSANNA HIETANEN

The largest pool of fixed nitrogen on Earth is dissolved organic nitrogen in the oceans, but its dynamics and transport between ocean zones are largely unknown. Writing in *Global Biogeochemical Cycles*, Letscher *et al.*[1] report that the observed concentration gradients of dissolved organic nitrogen in the oceans are the result of the interplay between high production in upwelling regions, the transfer of these waters from the surface to lower depths, and degradation by specifically adapted microbial communities. This contradicts the existing theory of the dynamics of nitrogen cycling.

The role of dissolved organic nitrogen (DON) in ocean productivity has long puzzled marine biogeochemists. DON accounts for roughly 60% of the reactive nitrogen in the ocean, which is a much higher percentage than that of readily available nitrogen-containing inorganic nutrients (such as nitrate) and particulate organic matter[2]. DON has therefore been proposed to be one of the major sources of nutrients in the open ocean, and to be responsible for the concentration gradients of nitrate that are observed at different regions and depths[3]. For a long time, DON was regarded as unsuitable for microbial uptake, but we now know that part of it fuels primary production — the formation

of organic molecules from carbon dioxide — especially in oceanic regions that are low in inorganic nutrients[4], such as open-ocean gyres (large systems of rotating currents).

However, we still lack a unifying theory for DON that combines global patterns of distribution and uptake with ocean productivity, seasonal variability and the behaviour of the microbial communities that metabolize its compounds. Letscher and colleagues' study goes some way towards solving this problem by bringing together the results of previously reported large-scale surveys of Atlantic DON concentrations[5,6] and modelling studies[7] with their own data derived from DON-degradation experiments, statistical models and observations from drifting buoys. They report that DON produced in eastern boundary upwelling areas located off the coasts of Mauritania, Namibia, Peru and Mexico is transported to subtropical gyres. But then what?

DON was generally assumed[6] to be mineralized (degraded) into inorganic nitrogen by microbes in surface waters (Fig. 1a), directly feeding export production — in which carbon is harvested by algal particles in surface waters, then transported to great depths as the dead algae sink. However, Letscher and co-workers' experiments revealed that DON is rather resistant to microbial degradation in surface waters, but is mineralized three times faster in the upper mesopelagic zone (waters at depths that receive some sunlight, but not enough for photosynthesis).

The authors therefore considered a scenario in which water mixing brings DON from the surface to the upper mesopelagic zone, where it is degraded; the subsequent transport of the resulting inorganic nutrients to the surface would then support export production (Fig. 1b). But when the researchers combined data from drifting buoys with statistical models, they found that this scenario was unlikely, because water-density differences prevent the DON-rich surface waters from reaching depths at which active microbial degradation occurs.

Letscher and colleagues thus suggest that a two-step process occurs. First, DON from surface water is transported to the lower euphotic zone — waters that lie immediately above the mesopelagic and receive sufficient sunlight for photosynthesis (Fig. 1c). Here, microbial communities break down DON into inorganic nitrogen that is taken up by algae. Second, the algae sink and are mineralized in the mesopelagic zone (Fig. 1c). The plausibility of this hypothesis largely depends on the capacity of microbial communities in the lower euphotic zone to break down DON compounds — something that remains to be investigated.

Despite this uncertainty, Letscher and colleagues' findings — that DON is mainly broken down below the ocean's surface layer and contributes significantly to the productivity of oceans that contain low levels of nutrients, especially at the eastern edges of the

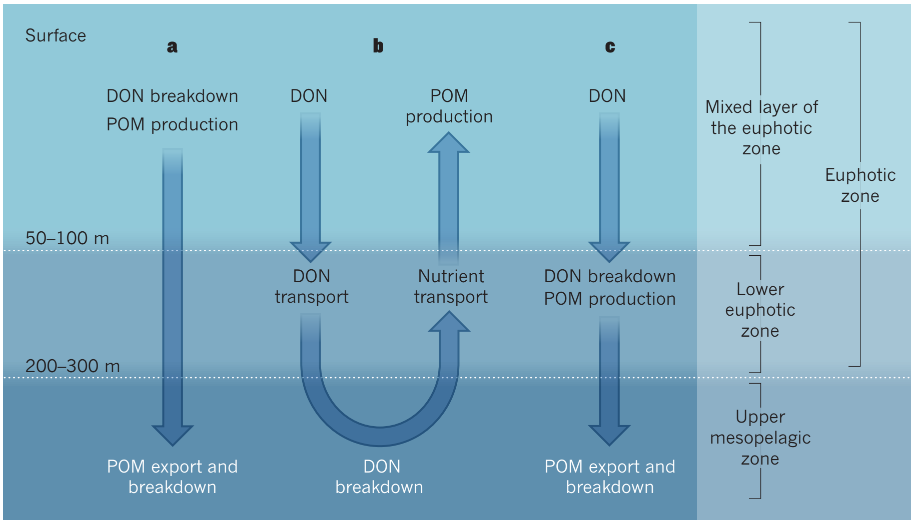


Figure 1 | Breaking down dissolved organic nitrogen (DON). **a**, In the conventional view of DON cycling in the ocean, DON is degraded by microbial communities in the mixed layer of the euphotic zone. The resulting inorganic nutrients support the photosynthetic harvesting of carbon by algae at the surface, which form particulate organic matter (POM) that sinks and transports the carbon to greater depths. Letscher *et al.*[1] report that this mechanism is unlikely, because surface microbes do not rapidly degrade DON. **b**, Alternatively, water mixing could transport DON to the mesopelagic zone, where it is broken down by local microbial communities. The resulting inorganic nutrients return to the surface, supporting POM production. However, the authors find that density differences prevent the transport of surface water to the mesopelagic. **c**, Instead, they suggest that DON is transported to, and degraded in, the lower euphotic zone, which receives enough light to support POM production. The POM then sinks to the mesopelagic zone, and is broken down.

subtropical gyres — are a great step forward in our understanding of marine DON dynamics, and fit well into present knowledge of global nitrogen cycling in the ocean. In subtropical gyres, mesoscale eddies (80 to 120 kilometres across) provide only approximately one-third of the nitrate that is needed to explain the observed productivity[8]. DON might be another contributing nitrogen source.

The biggest remaining challenges to our understanding of DON dynamics are working out how available DON is to microbes and determining its molecular variability. The recognition of a distinction between high- and low-molecular-weight fractions of DON has provided some insight into DON cycling. The isotopic composition of nitrogen in sampled high-molecular-weight fractions shows little temporal or spatial variation, even though nitrate concentrations and the nitrogen-isotope composition of suspended particulate organic matter taken from the same locations fluctuate. This suggests that most DON is cycled on timescales of less than a year[9]. High-molecular-weight fractions at the surface seem to be more readily mineralized than those in deeper zones[10]. Carbon dating of these fractions also shows that they are more reactive than the low-molecular-weight fractions[11].

However, studies of the two fractions have provided only the average activities of the diverse mixtures of compounds within them. Fortunately, ultra-high-resolution mass spectrometry can now simultaneously identify thousands of different molecules within dissolved organic matter[12]. This reveals the composition of dissolved organic matter in unprecedented detail — not only for the fraction that is DON, but also for soluble forms of phosphorus that are relevant to nitrogen-fixing organisms, and carbon used by heterotrophic bacteria for growth. Attention should therefore focus on bacterial communities and their abilities to degrade various organic compounds.

In the near future, thawing permafrost in the Arctic tundra might release large quantities of DON of unknown composition[13], which could fuel the productivity of the ocean at adjacent coasts. Along with increasing loads of nitrate from the atmosphere and from rivers, this would add yet more reactive nitrogen to the oceans. The extent to which these sources will enhance productivity is unknown, but they could be crucial for all of the oceans. Studies such as those of Letscher *et al.* might help us to address this issue. ■

Maren Voss *is in the Department of Biological Oceanography, Leibniz Institute for Baltic Sea Research, Warnemünde, 18119 Rostock, Germany.* **Susanna Hietanen** *is in the Department of Environmental Sciences, University of Helsinki, 00014 Helsinki, Finland.*
e-mails: maren.voss@io-warnemuende.de; susanna.hietanen@helsinki.fi

1. Letscher, R. T., Hansell, D. A., Carlson, C. A., Lumpkin, R. & Knapp, A. N. *Glob. Biogeochem. Cycles* http://dx.doi.org/10.1029/2012GB004449 (2012).
2. Bronk, D. A. in *Biogeochemistry of Marine Dissolved Organic Matter* (eds Hansell, D. A. & Carlson, C. A.) 153–200 (Academic, 2002).

3. Charria, G., Dadou, I., Llido, J., Drevillon, M. & Garcon, V. *Biogeosciences* **5,** 1437–1455 (2008).
4. Bronk, D. A., See, J. H., Bradley, P. & Killberg, L. *Biogeosciences* **4,** 283–296 (2007).
5. Mahaffey, C., Williams, R. G. & Wolff, G. A. *Glob. Biogeochem. Cycles* **18,** GB1034 (2004).
6. Torres-Valdés, S. *et al. Glob. Biogeochem. Cycles* **23,** GB4019 (2009).
7. Roussenov, V., Williams, R. G., Mahaffey, C. & Wolff, G. A. *Glob. Biogeochem. Cycles* **20,** GB3002 (2006).
8. Oschlies, A. & Garcon, V. *Nature* **394,** 266–269 (1998).
9. Meador, T. B., Aluwihare, L. I. & Mahaffey, C. *Limnol. Oceanogr.* **52,** 934–947 (2007).
10. Aluwihare, L. I., Repeta, D. J., Pantoja, S. & Johnson, C. G. *Science* **308,** 1007–1010 (2005).
11. Loh, A. N., Bauer, J. E. & Druffel, E. R. M. *Nature* **430,** 877–881 (2004).
12. Koch, B. P., Witt, M., Engbrodt, R., Dittmar, T. & Kattner, G. *Geochim. Cosmochim. Acta* **69,** 3299–3308 (2005).
13. Frey, K. E., McClelland, J. W., Holmes, R. M. & Smith, L. C. *J. Geophys. Res.* **112,** G04S58 (2007).

MATERIALS SCIENCE

Synthetic polymers with biological rigidity

Brush-like polymers with a rigidity similar to that of polymers in living cells have been synthesized and used to build force-responsive materials. The advance opens the door to applications in drug delivery and tissue engineering. SEE LETTER P.651

MARGARET LISE GARDEL

The diverse physiology of cells and tissues is underpinned by materials consisting of macromolecules whose mechanical behaviour allows organisms to control and maintain their shape[1]. The construction of synthetic versions of these materials could allow artificial cells and tissues to be made, but preparing such materials has long been a major challenge. On page 651 of this issue, Kouwer *et al.*[2] report that they have produced the first synthetic polymers whose rigidity can be tuned to mimic that of a wide range of their biological counterparts. The authors' achievement will facilitate the construction of polymer networks that have highly tunable, force-responsive behaviour*.

Synthetic polymers, such as polyethylene, nylon and silicone, were an important class of material in the twentieth century, finding diverse applications as paints, adhesives, fibres and plastics. But these polymer molecules behave rather like cooked spaghetti, because they have little rigidity along their length. Their flexibility is entirely due to the randomization of the polymer chains' configurations by thermal energy — the energy available to act on molecules at ambient temperature.

Biological polymers, which are formed from amino acids or nucleic acids, are very different. These materials are ubiquitous in nature, and include DNA; cytoskeletal filamentous proteins, such as actin, microtubules and intermediate filaments; and scaffolding molecules in the extracellular matrix, such as collagen and fibrin. Biological polymers are much more rigid than chemical polymers, and so are similar to partially cooked spaghetti. Because of this high rigidity, the energy required to bend biological polymers is comparable to that available from thermal energy, such that they bend much less than synthetic polymers at ambient temperature. This inherent rigidity makes the mechanical behaviour of biological polymers at the bulk scale qualitatively different from that of synthetic polymers[3].

Kouwer and colleagues have discovered that polyisocyanopeptide polymers, grafted with flexible side chains of a different polymer, serve as mimics of a protein structure known as a β-sheet, and self-assemble into helical structures similar to those formed by DNA and actin filaments. Moreover, the authors report that the polymers aggregate into bundles when heated in solution (Fig. 1), similar to the bundles formed by collagen and fibrin.

One way to characterize the rigidity of a material is through its persistence length: the higher the persistence length, the more rigid the polymer. The persistence length of biological polymers varies from about 100 nanometres for DNA to 1 millimetre for microtubules; by comparison, the effective persistence length of a flexible synthetic polymer is typically about 0.1 nm. When the authors characterized the mechanical properties of their polymers using force-spectroscopy techniques, they found that single polymer chains had a sizeable persistence length, 500 nm. They also found that this increased for larger bundles, consistent with the idea that rigidity correlates with bundle diameter. Kouwer and co-workers' materials therefore represent the first semi-flexible synthetic polymers to have tunable persistence lengths, and so might serve as building blocks for biomimetic materials.

One of the main consequences of increasing polymer rigidity is that it alters the mechanical response of crosslinked networks of the polymer. The mechanical rigidity of a material is described by a parameter known as the elastic modulus. For networks of flexible polymers (such as rubber), the elastic modulus depends only weakly on the density of the polymers or crosslinker connections. By contrast, in networks of semi-flexible polymers, the elastic modulus depends more strongly on these parameters[3].

A second characteristic of networks of semi-flexible polymers is that their response to stress is highly nonlinear[3]. Under increasing loads, conventional polymeric materials simply stretch until they break. Networks of semi-flexible polymers, however, stiffen under increasing load and have an elastic modulus that increases dramatically at a critical strain. Both of these distinctive properties of semi-flexible networks are typical of biological polymers, and are also found in Kouwer and colleagues' synthetic polymers.

Because of their unusual mechanics, the authors' materials are highly responsive to applied stress: as the applied stress increases,

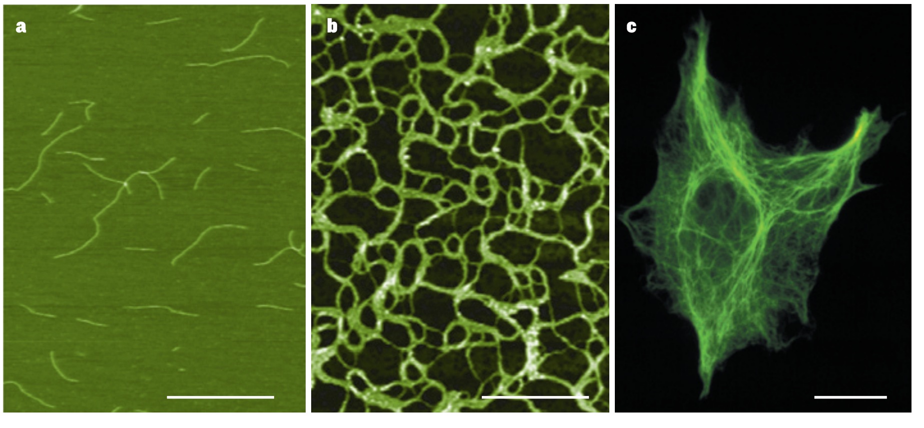


Figure 1 | Bundling fibres. Kouwer *et al.*[2] report the first synthetic polymer that has rigidity similar to that of biological polymers such as DNA. Single chains of the polymers (**a**) form bundles (**b**) when heated in solution. The polymers most resemble those found in intermediate filaments (**c**) inside cells. Scale bars: **a**, **b**, 250 nm; **c**, 85 μm.

*This article and the paper under discussion[2] were published online on 23 January 2013.

3. Charria, G., Dadou, I., Llido, J., Drevillon, M. & Garcon, V. *Biogeosciences* **5,** 1437–1455 (2008).
4. Bronk, D. A., See, J. H., Bradley, P. & Killberg, L. *Biogeosciences* **4,** 283–296 (2007).
5. Mahaffey, C., Williams, R. G. & Wolff, G. A. *Glob. Biogeochem. Cycles* **18,** GB1034 (2004).
6. Torres-Valdés, S. *et al. Glob. Biogeochem. Cycles* **23,** GB4019 (2009).
7. Roussenov, V., Williams, R. G., Mahaffey, C. & Wolff, G. A. *Glob. Biogeochem. Cycles* **20,** GB3002 (2006).
8. Oschlies, A. & Garcon, V. *Nature* **394,** 266–269 (1998).
9. Meador, T. B., Aluwihare, L. I. & Mahaffey, C. *Limnol. Oceanogr.* **52,** 934–947 (2007).
10. Aluwihare, L. I., Repeta, D. J., Pantoja, S. & Johnson, C. G. *Science* **308,** 1007–1010 (2005).
11. Loh, A. N., Bauer, J. E. & Druffel, E. R. M. *Nature* **430,** 877–881 (2004).
12. Koch, B. P., Witt, M., Engbrodt, R., Dittmar, T. & Kattner, G. *Geochim. Cosmochim. Acta* **69,** 3299–3308 (2005).
13. Frey, K. E., McClelland, J. W., Holmes, R. M. & Smith, L. C. *J. Geophys. Res.* **112,** G04S58 (2007).

MATERIALS SCIENCE

Synthetic polymers with biological rigidity

Brush-like polymers with a rigidity similar to that of polymers in living cells have been synthesized and used to build force-responsive materials. The advance opens the door to applications in drug delivery and tissue engineering. SEE LETTER P.651

MARGARET LISE GARDEL

The diverse physiology of cells and tissues is underpinned by materials consisting of macromolecules whose mechanical behaviour allows organisms to control and maintain their shape[1]. The construction of synthetic versions of these materials could allow artificial cells and tissues to be made, but preparing such materials has long been a major challenge. On page 651 of this issue, Kouwer *et al.*[2] report that they have produced the first synthetic polymers whose rigidity can be tuned to mimic that of a wide range of their biological counterparts. The authors' achievement will facilitate the construction of polymer networks that have highly tunable, force-responsive behaviour*.

Synthetic polymers, such as polyethylene, nylon and silicone, were an important class of material in the twentieth century, finding diverse applications as paints, adhesives, fibres and plastics. But these polymer molecules behave rather like cooked spaghetti, because they have little rigidity along their length. Their flexibility is entirely due to the randomization of the polymer chains' configurations by thermal energy — the energy available to act on molecules at ambient temperature.

Biological polymers, which are formed from amino acids or nucleic acids, are very different. These materials are ubiquitous in nature, and include DNA; cytoskeletal filamentous proteins, such as actin, microtubules and intermediate filaments; and scaffolding molecules in the extracellular matrix, such as collagen and fibrin. Biological polymers are much more rigid than chemical polymers, and so are similar to partially cooked spaghetti. Because of this high rigidity, the energy required to bend biological polymers is comparable to that available from thermal energy, such that they bend much less than synthetic polymers at ambient temperature. This inherent rigidity makes the mechanical behaviour of biological polymers at the bulk scale qualitatively different from that of synthetic polymers[3].

Kouwer and colleagues have discovered that polyisocyanopeptide polymers, grafted with flexible side chains of a different polymer, serve as mimics of a protein structure known as a β-sheet, and self-assemble into helical structures similar to those formed by DNA and actin filaments. Moreover, the authors report that the polymers aggregate into bundles when heated in solution (Fig. 1), similar to the bundles formed by collagen and fibrin.

One way to characterize the rigidity of a material is through its persistence length: the higher the persistence length, the more rigid the polymer. The persistence length of biological polymers varies from about 100 nanometres for DNA to 1 millimetre for microtubules; by comparison, the effective persistence length of a flexible synthetic polymer is typically about 0.1 nm. When the authors characterized the mechanical properties of their polymers using force-spectroscopy techniques, they found that single polymer chains had a sizeable persistence length, 500 nm. They also found that this increased for larger bundles, consistent with the idea that rigidity correlates with bundle diameter. Kouwer and co-workers' materials therefore represent the first semi-flexible synthetic polymers to have tunable persistence lengths, and so might serve as building blocks for biomimetic materials.

One of the main consequences of increasing polymer rigidity is that it alters the mechanical response of crosslinked networks of the polymer. The mechanical rigidity of a material is described by a parameter known as the elastic modulus. For networks of flexible polymers (such as rubber), the elastic modulus depends only weakly on the density of the polymers or crosslinker connections. By contrast, in networks of semi-flexible polymers, the elastic modulus depends more strongly on these parameters[3].

A second characteristic of networks of semi-flexible polymers is that their response to stress is highly nonlinear[3]. Under increasing loads, conventional polymeric materials simply stretch until they break. Networks of semi-flexible polymers, however, stiffen under increasing load and have an elastic modulus that increases dramatically at a critical strain. Both of these distinctive properties of semi-flexible networks are typical of biological polymers, and are also found in Kouwer and colleagues' synthetic polymers.

Because of their unusual mechanics, the authors' materials are highly responsive to applied stress: as the applied stress increases,

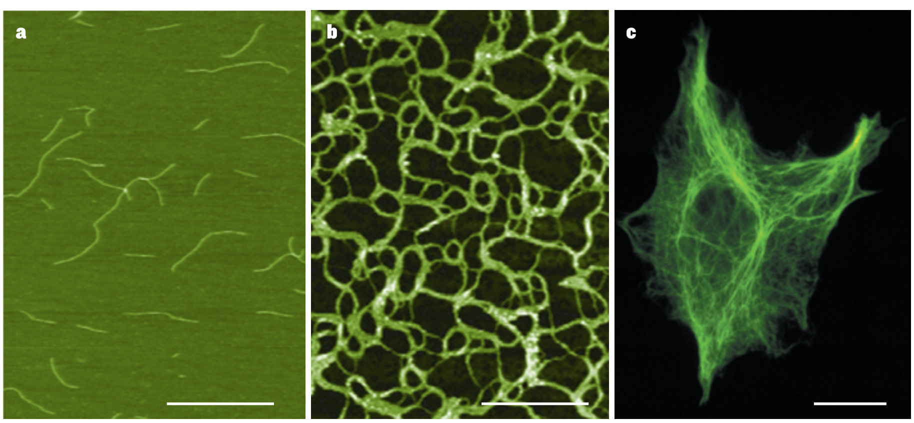


Figure 1 | Bundling fibres. Kouwer *et al.*[2] report the first synthetic polymer that has rigidity similar to that of biological polymers such as DNA. Single chains of the polymers (**a**) form bundles (**b**) when heated in solution. The polymers most resemble those found in intermediate filaments (**c**) inside cells. Scale bars: **a**, **b**, 250 nm; **c**, 85 μm.

*This article and the paper under discussion[2] were published online on 23 January 2013.

the elastic modulus also increases to minimize changes in deformation. This suggests that the materials will maintain their shape when subjected to a wide range of externally applied stresses. What's more, the polymers' highly tunable rigidity means that tiny quantities of polymer could be used to make materials that have a wide range of stiffnesses.

Kouwer and colleagues' polymers most closely mimic those found in intermediate filaments (Fig. 1), a class of intracellular polymer that is crucial for cell adhesion and migration and for maintaining cell shape[4]. It will be exciting to see if the authors' approach, or other approaches for making semi-flexible polymers, can be expanded to make synthetic mimics of DNA, actin filaments and microtubules. Another challenge will be to find a way of adding mechanochemically active components[5] — those that transform chemical energy into mechanical work — to the polymer. This would enable filaments to be made that exhibit exotic polymerization behaviour, such as treadmilling (in which one end of a filament grows while its other end shrinks), or which create dynamic instabilities or crosslinks, to form the basis of a molecular motor.

The ability to build 'active' soft materials that respond to external chemical and mechanical signals will provide opportunities in the areas of condensed-matter physics and materials science for years to come. Such materials might allow the construction of artificial cells and tissues that are more closely compatible physiologically with their counterparts in humans than currently available materials, so that they might be used in the next generation of drug-delivery and tissue-engineering technologies. Active soft materials might also change the way in which we engage with the physical world, by forming the basis of highly responsive and malleable materials and machines. Kouwer and co-workers' polymers are an exciting first step in these directions. ■

➲ NATURE.COM
For more on synthetic polymers, see:
go.nature.com/xbhltj

Margaret Lise Gardel *is at the Gordon Center for Integrated Science, Department of Physics, University of Chicago, Illinois 60637, USA.*
e-mail: gardel@uchicago.edu

1. Fletcher, D. A. & Mullins, R. D. *Nature* **463,** 485–492 (2010).
2. Kouwer, P. H. J. *et al. Nature* **493,** 651–655 (2013).
3. Gardel, M. L. *et al. Meth. Cell Biol.* **89,** 487–519 (2008).
4. Goldman, R. D. *et al. J. Struct. Biol.* **177,** 14–23 (2012).
5. Fletcher, D. A. & Geissler, P. L. *Annu. Rev. Phys. Chem.* **60,** 469–486 (2009).

CONDENSED-MATTER PHYSICS

Hidden is more

Physicists have puzzled over a hidden electronic order in a uranium-based material for decades. A new theory attributes it to not just a single but a double breaking of time-reversal symmetry. SEE ARTICLE P.621

QIMIAO SI

A magnet sticks to a fridge door, but an aluminium spoon does not. This distinction is well understood in terms of the different ways in which the many billions of billions of electrons are collectively organized inside these materials. In a magnet, the electrons form an order: their tiny spins line up along a particular direction, producing an aggregate magnetic moment, whereas in aluminium these spins are randomly oriented. On page 621 of this issue, Chandra *et al.*[1] propose a different kind of electronic order, which could resolve a riddle that has confounded physicists for more than a quarter of a century.

Much of the fascination and challenge of condensed-matter physics lies in figuring out how the electrons are organized in their microscopic world to produce the macroscopic properties observed in the laboratory. The tendency of the electrons in a magnet to develop order is analogous to that of water molecules to form a rigid spatial pattern as the liquid freezes into ice. This electronic order, called ferromagnetism, breaks time-reversal symmetry: if the time direction were reversed, so would be the direction of the magnetic moment.

The condensed-matter system studied by Chandra and colleagues is URu_2Si_2. This uranium-based compound is a member of a broad class of materials called strongly correlated electron systems, in which a large Coulomb repulsion between the electrons tends to produce spectacular physical phenomena — such as the high-temperature superconductivity observed in copper-based ceramics. This large repulsion in strongly correlated electron systems contrasts with the weak interactions found in many of the materials used in technology, such as silicon, aluminium or even ordinary magnets.

In the mid-1980s, researchers discovered[2–4] clear signatures of an electronic order in

NATURE

50 Years Ago

Living with the Atom. By Prof. Ritchie Calder — The author gives his … contributions to a discussion on responsible reporting. It is difficult for the reporter to steer a course between the attractive liveliness that approaches the sensational, and the dry factual report that few will read … There is a fear and distrust of scientists, as people who ought to be above human fallibility but unforgivably err like everyone else; and he points out the dangers also of the responsibility for major decisions resting in the hands of men who do not know sufficient about science to be able to challenge with any confidence the advice that comes to them from their experts. There is something in this, although one would feel that a knowledge of men and a flair for consulting the right experts is what brings men to high office.
From *Nature* 2 February 1963

100 Years Ago

'Luminous halos surrounding shadows of heads' — The phenomenon referred to … can also be seen on grass when the sun is low in the sky … If the grass surface is near to the observer, a faint halo is seen to surround the shadow of his head, and this is more easily perceived if he is moving than if standing still; my attention was indeed first attracted to this phenomenon when bicycling. J. Evershed
I happened to be watching our shadows as we passed along the edge of a field of young green wheat, when, to my surprise, I noticed a halo of light round the shadow of my own head and neck … The fact that each observer sees only his own halo obviously precludes this phenomenon from having been the origin of the halos recorded in sacred writings round the head of Christ and others. L. L. Fermor
From *Nature* 30 January 1913

the elastic modulus also increases to minimize changes in deformation. This suggests that the materials will maintain their shape when subjected to a wide range of externally applied stresses. What's more, the polymers' highly tunable rigidity means that tiny quantities of polymer could be used to make materials that have a wide range of stiffnesses.

Kouwer and colleagues' polymers most closely mimic those found in intermediate filaments (Fig. 1), a class of intracellular polymer that is crucial for cell adhesion and migration and for maintaining cell shape[4]. It will be exciting to see if the authors' approach, or other approaches for making semi-flexible polymers, can be expanded to make synthetic mimics of DNA, actin filaments and microtubules. Another challenge will be to find a way of adding mechanochemically active components[5] — those that transform chemical energy into mechanical work — to the polymer. This would enable filaments to be made that exhibit exotic polymerization behaviour, such as treadmilling (in which one end of a filament grows while its other end shrinks), or which create dynamic instabilities or crosslinks, to form the basis of a molecular motor.

➲ NATURE.COM
For more on synthetic polymers, see:
go.nature.com/xbhltj

The ability to build 'active' soft materials that respond to external chemical and mechanical signals will provide opportunities in the areas of condensed-matter physics and materials science for years to come. Such materials might allow the construction of artificial cells and tissues that are more closely compatible physiologically with their counterparts in humans than currently available materials, so that they might be used in the next generation of drug-delivery and tissue-engineering technologies. Active soft materials might also change the way in which we engage with the physical world, by forming the basis of highly responsive and malleable materials and machines. Kouwer and co-workers' polymers are an exciting first step in these directions. ■

Margaret Lise Gardel *is at the Gordon Center for Integrated Science, Department of Physics, University of Chicago, Illinois 60637, USA.*
e-mail: gardel@uchicago.edu

1. Fletcher, D. A. & Mullins, R. D. *Nature* **463,** 485–492 (2010).
2. Kouwer, P. H. J. *et al. Nature* **493,** 651–655 (2013).
3. Gardel, M. L. *et al. Meth. Cell Biol.* **89,** 487–519 (2008).
4. Goldman, R. D. *et al. J. Struct. Biol.* **177,** 14–23 (2012).
5. Fletcher, D. A. & Geissler, P. L. *Annu. Rev. Phys. Chem.* **60,** 469–486 (2009).

CONDENSED-MATTER PHYSICS

Hidden is more

Physicists have puzzled over a hidden electronic order in a uranium-based material for decades. A new theory attributes it to not just a single but a double breaking of time-reversal symmetry. SEE ARTICLE P.621

QIMIAO SI

A magnet sticks to a fridge door, but an aluminium spoon does not. This distinction is well understood in terms of the different ways in which the many billions of billions of electrons are collectively organized inside these materials. In a magnet, the electrons form an order: their tiny spins line up along a particular direction, producing an aggregate magnetic moment, whereas in aluminium these spins are randomly oriented. On page 621 of this issue, Chandra *et al.*[1] propose a different kind of electronic order, which could resolve a riddle that has confounded physicists for more than a quarter of a century.

Much of the fascination and challenge of condensed-matter physics lies in figuring out how the electrons are organized in their microscopic world to produce the macroscopic properties observed in the laboratory. The tendency of the electrons in a magnet to develop order is analogous to that of water molecules to form a rigid spatial pattern as the liquid freezes into ice. This electronic order, called ferromagnetism, breaks time-reversal symmetry: if the time direction were reversed, so would be the direction of the magnetic moment.

The condensed-matter system studied by Chandra and colleagues is URu_2Si_2. This uranium-based compound is a member of a broad class of materials called strongly correlated electron systems, in which a large Coulomb repulsion between the electrons tends to produce spectacular physical phenomena — such as the high-temperature superconductivity observed in copper-based ceramics. This large repulsion in strongly correlated electron systems contrasts with the weak interactions found in many of the materials used in technology, such as silicon, aluminium or even ordinary magnets.

In the mid-1980s, researchers discovered[2–4] clear signatures of an electronic order in

50 Years Ago

Living with the Atom. By Prof. Ritchie Calder — The author gives his … contributions to a discussion on responsible reporting. It is difficult for the reporter to steer a course between the attractive liveliness that approaches the sensational, and the dry factual report that few will read … There is a fear and distrust of scientists, as people who ought to be above human fallibility but unforgivably err like everyone else; and he points out the dangers also of the responsibility for major decisions resting in the hands of men who do not know sufficient about science to be able to challenge with any confidence the advice that comes to them from their experts. There is something in this, although one would feel that a knowledge of men and a flair for consulting the right experts is what brings men to high office.
From *Nature* 2 February 1963

100 Years Ago

'Luminous halos surrounding shadows of heads' — The phenomenon referred to … can also be seen on grass when the sun is low in the sky … If the grass surface is near to the observer, a faint halo is seen to surround the shadow of his head, and this is more easily perceived if he is moving than if standing still; my attention was indeed first attracted to this phenomenon when bicycling. J. Evershed
I happened to be watching our shadows as we passed along the edge of a field of young green wheat, when, to my surprise, I noticed a halo of light round the shadow of my own head and neck … The fact that each observer sees only his own halo obviously precludes this phenomenon from having been the origin of the halos recorded in sacred writings round the head of Christ and others. L. L. Fermor
From *Nature* 30 January 1913

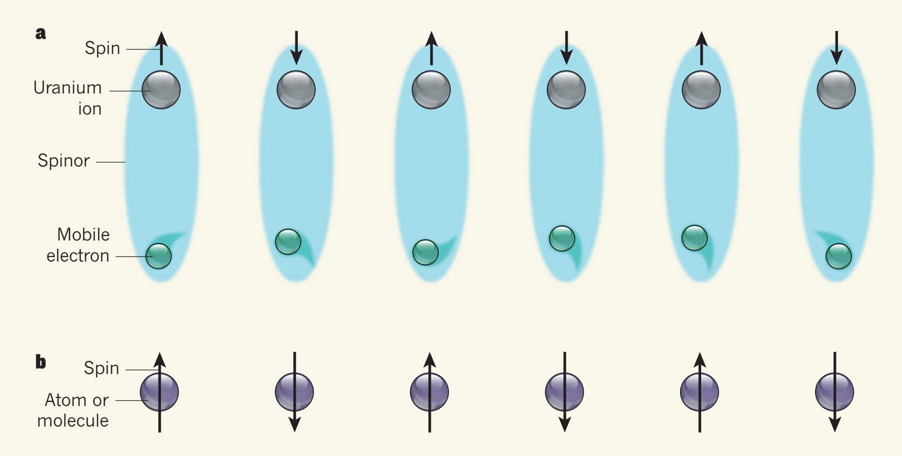


Figure 1 | Magnetic orders. **a**, The spinor order in URu_2Si_2 proposed by Chandra and colleagues[1]. The uranium ions have two electrons in their outermost 5*f* orbitals. Together, the electrons have an angular momentum of $\hbar$, where $\hbar$ is Planck's reduced constant. Mobile electrons in the material have an angular momentum of half of $\hbar$ and 'hybridize' with the uranium ions, producing entities called spinors that carry a spin of half $\hbar$. The spinors form an arrangement of spins that align antiparallel to each other. The authors propose that this form of order breaks symmetry in a single time-reversal transformation as well as in a double time reversal. **b**, Ordinary antiferromagnetic order. The electron spins of atoms or molecules carry an angular momentum that is an integer of $\hbar$. The adjacent spins point in opposite directions, generating an order that breaks symmetry in a single but not a double time reversal.

URu_2Si_2 when the material was cooled below 17.5 kelvin. But the order was mysterious: it was different from ferromagnetism, antiferromagnetism (Fig. 1) or any other order known in the magnetic world. Since then, more than two dozen theoretical ideas[5] have been put forward as candidate orders for URu_2Si_2. Some have been invalidated by experiments, whereas others remain a matter of contention. Condensed-matter physicists, in frustration, have referred to the phenomenon as a hidden order.

To make progress, Chandra *et al.* went back to basics. The spin of an electron has its origin in quantum mechanics, which divides subatomic particles into two categories: bosons and fermions. Electrons are fermions and, unlike bosons, cannot share the same quantum state. Nonetheless, they can quantum-mechanically entangle with each other. This entanglement is deeply ingrained in our understanding of heavy-fermion metals[6], which make up a prominent family of materials within the strongly correlated electron systems to which URu_2Si_2 belongs.

The entanglement of itinerant (mobile) electrons with strongly correlated electrons that are localized on the uranium ions of URu_2Si_2 inhibits the motion of the itinerant electrons, and effectively enhances their mass by a huge factor — typically in the hundreds — compared with the bare-electron mass. The entanglement also mixes up the identities of the itinerant and localized electrons, a process called hybridization. The spins of the electrons in such a hybridized state can point in any direction, and no symmetry is broken.

Chandra and colleagues examined the details of this hybridization in URu_2Si_2, a crystal comprising layers of atomic planes. The process involves quantum tunnelling of itinerant electrons into or out of the compound's uranium ions, resulting in an odd number of electrons in the ion's 5*f* orbitals and two excited electronic states of opposite spin orientation on each ion. The two states, known as a Kramers doublet, are connected to each other by a time-reversal transformation.

On theoretical grounds, Chandra *et al.* have proposed that lowering the temperature of the material induces an order in the hybridization that breaks the time-reversal symmetry. More precisely, unlike in ordinary magnets — in which the elementary unit of the order is a spin carrying an angular momentum that is an integer amount of Planck's reduced constant ($\hbar$) — in the proposed order, the elementary unit has an angular momentum of one-half of $\hbar$. Consequently, when a time-reversal operation is applied twice to the system, symmetry is not restored. In other words, the order breaks symmetry not only in a single time reversal, but also in a double time reversal. The elementary unit of the proposed order forms a mathematical object known as a spinor (Fig. 1).

On the basis of this theoretical proposal, Chandra *et al.* have provided an explanation for several of the intriguing properties observed in URu_2Si_2, including a striking magnetic anisotropy[7] — a large difference between the system's responses to a magnetic field that is applied parallel to the atomic planes, and to one that is applied perpendicular to the planes. The theory also includes an earlier-derived feature that connects the hidden-order state with a pressure-induced antiferromagnetic state[8]. These results make the proposed order a leading contender for the hidden order in URu_2Si_2. However, the theory rests on specific assumptions about the electronic configurations of the 5*f* orbitals in the uranium ions that should be tested experimentally. The evolution of the hybridization process as the temperature is lowered, from one that preserves all symmetries to one that breaks time-reversal invariance, should also be investigated by measuring the momentum dependence of the electronic states using photoemission spectroscopy[9] and electronic tunnelling spectroscopy[10–12].

From a theoretical perspective, the proposed spinor order is a refreshing idea. Ordinarily, whereas the onset of hybridization at absolute-zero temperature represents a sharp phase transition[6], its thermally induced counterpart is only a gradual crossover phenomenon. Because the proposed spinor order breaks time-reversal symmetry, it turns the hybridization onset into a sharp phase transition even at a non-zero temperature. However, a spinor is usually a fermionic object and therefore is not allowed to order. Chandra *et al.* introduced an approximate treatment of the strong electron-correlation effects that made the spinor order possible, but this theoretical procedure requires further elucidation.

Regardless of what future investigations may uncover, Chandra and colleagues' study opens up a new dimension in the ongoing debate about the nature of the hidden order in URu_2Si_2, and enriches our exploration of strongly correlated matter that breaks time-reversal symmetry. More generally, strong correlations often produce competing tendencies for electronic order, which, in turn, foster the emergence of electronic phenomena such as superconductivity. Hence, the spinor order proposed here, as well as other exotic orders in related strongly correlated materials, will offer insight into collective electronic organization that may lead us to understand pressing issues such as high-temperature superconductivity. ■

Qimiao Si *is in the Department of Physics and Astronomy, Rice University, Houston, Texas 77005, USA.*
e-mail: qmsi@rice.edu

1. Chandra, P., Coleman, P. & Flint, R. *Nature* **493,** 621–626 (2013).
2. Palstra, T. T. M. *et al. Phys. Rev. Lett.* **55,** 2727–2730 (1985).
3. Schlabitz, W. *et al. Z. Phys.* **B62,** 171–177 (1986).
4. Maple, M. B. *et al. Phys. Rev. Lett.* **56,** 185–188 (1986).
5. Mydosh, J. A. & Oppeneer, P. M. *Rev. Mod. Phys.* **83,** 1301–1322 (2011).
6. Si, Q. & Steglich, F. *Science* **329,** 1161–1166 (2010).
7. Altarawneh, M. M. *et al. Phys. Rev. Lett.* **106,** 146403 (2011).
8. Haule, K. & Kotliar, G. *Europhys. Lett.* **89,** 57006 (2010).
9. Santander-Syro, A. F. *et al. Nature Phys.* **5,** 637–641 (2009).
10. Schmidt, A. R. *et al. Nature* **465,** 570–576 (2010).
11. Aynajian, P. *et al. Proc. Natl Acad. Sci. USA* **107,** 10383–10388 (2010).
12. Park, W. K. *et al. Phys. Rev. Lett.* **108,** 246403 (2012).

doi:10.1038/nature11820

Hastatic order in the heavy-fermion compound URu_2Si_2

Premala Chandra[1], Piers Coleman[1,2] & Rebecca Flint[3]

The development of collective long-range order by means of phase transitions occurs by the spontaneous breaking of fundamental symmetries. Magnetism is a consequence of broken time-reversal symmetry, whereas superfluidity results from broken gauge invariance. The broken symmetry that develops below 17.5 kelvin in the heavy-fermion compound URu_2Si_2 has long eluded such identification. Here we show that the recent observation of Ising quasiparticles in URu_2Si_2 results from a spinor order parameter that breaks double time-reversal symmetry, mixing states of integer and half-integer spin. Such 'hastatic' order hybridizes uranium-atom conduction electrons with Ising $5f^2$ states to produce Ising quasiparticles; it accounts for the large entropy of condensation and the magnetic anomaly observed in torque magnetometry. Hastatic order predicts a tiny transverse moment in the conduction-electron 'sea', a colossal Ising anisotropy in the nonlinear susceptibility anomaly and a resonant, energy-dependent nematicity in the tunnelling density of states.

The hidden order that develops below $T_{HO} = 17.5$ K in the heavy-fermion compound URu_2Si_2 is particularly notable, having eluded identification for 25 years[1–12]. Recent spectroscopic[13–17], magnetometric[18] and high-field measurements[19,20] suggest that the hidden order is connected with the formation of an itinerant heavy-electron fluid, as a consequence of quasiparticle hybridization between localized, spin–orbit-coupled f-shell moments and mobile conduction electrons. Although the development of hybridization at low temperatures is usually associated with a crossover, in URu_2Si_2 both optical[17] and tunnelling[14–16] probes suggest that it develops abruptly at the hidden-order transition, leading to proposals[9,10] that the hybridization is an order parameter.

Ising quasiparticles

High-temperature bulk susceptibility measurements on URu_2Si_2 show that the local $5f$ moments embedded in the conduction-electron sea are Ising in nature[1,21], and quantum oscillation experiments deep within the hidden-order phase[22] reveal that the quasiparticles possess a giant Ising anisotropy[20,23,24]. The Zeeman splitting $\Delta E(\theta)$ depends solely on the c-axis component of the magnetic field: $\Delta E = g(\theta)\mu_B B$ (ref. 24). Here B is the magnetic field, μ_B is the Bohr magneton and the empirically determined g-factor takes the form $g(\theta) = g\cos(\theta)$, where θ is the angle between the magnetic field and the c axis and g is the Ising g-factor. The g-factor anisotropy exceeds 30, corresponding to an anisotropy of the Pauli susceptibility in excess of 900; this anisotropy is also observed in the angle dependence of the Pauli-limited upper critical field of the superconducting state[23,24], showing that the Ising quasiparticles pair to form a heavy-fermion superconductor. This giant anisotropy suggests that the f moment is transferred to the mobile quasiparticles through hybridization[25].

In the tetragonal crystalline environment of URu_2Si_2, such Ising anisotropy is most natural in an integer-spin $5f^2$ configuration of the uranium ions[4,26]. Although a variety of singlet crystal-field schemes have been proposed[6,27], the observation of paired Ising quasiparticles in a superconductor with a transition temperature of $T_c \approx 1.5$ K indicates that this $5f^2$ configuration is doubly degenerate to within an energy resolution of $g\mu_B H_{c2} \approx 5$ K, where H_{c2} is the upper critical field of the superconductor. Moreover, the observation of multiple spin zeroes in the quantum oscillations, resulting from the interference of Zeeman split orbits in a tilted field, requires that in a transverse field the underlying $5f^2$ configuration is doubly degenerate to within a cyclotron energy, which is $\hbar\omega_c = \hbar eB/m^* \approx 1.5$ K for the largest extremal orbit[20,22] (α) ($m^* = 12.5 m_e$ measured in $B = 13.9$ T, where m_e is the electron mass). These tiny bounds suggest that the Ising $5f^2$ state is intrinsically degenerate. In URu_2Si_2, tetragonal symmetry protects such a magnetic non-Kramers Γ_5 doublet[28], the candidate origin of the Ising quasiparticles[4,29].

The quasiparticle hybridization of half-integer-spin conduction electrons with an integer-spin doublet in URu_2Si_2 has profound implications for hidden order; such mixing can not occur without the breaking of double time-reversal symmetry. Time-reversal, $\hat{\Theta}$, is an anti-unitary quantum operator with no associated quantum number[30]. However double time-reversal, $\hat{\Theta}^2$, which is equivalent to a 2π rotation, forms a unitary operator with an associated quantum number, the 'Kramers index', K (ref. 30). For a quantum state of total angular momentum J, $K = (-1)^{2J}$ defines the phase factor acquired by its wavefunction after two successive time-reversals: $\hat{\Theta}^2|\psi\rangle = K|\psi\rangle = |\psi^{2\pi}\rangle$. An integer-spin state $|\alpha\rangle$ is unchanged by a 2π rotation, and so $|\alpha^{2\pi}\rangle = +|\alpha\rangle$ and $K = 1$. However, conduction electrons with half-integer-spin states, $|k\sigma\rangle$, where k is the vector momentum and σ is the spin component, change sign: $|k\sigma^{2\pi}\rangle = -|k\sigma\rangle$. Hence, $K = -1$ for conduction electrons.

Double time-reversal symmetry

Although conventional magnetism breaks time-reversal symmetry, it is invariant under $\hat{\Theta}^2$, with the result that the Kramers index is conserved. However, in URu_2Si_2 the hybridization between integer-spin and half-integer-spin states requires a quasiparticle mixing term of the form $\mathcal{H} = |k\sigma\rangle V_{\sigma\alpha}(k)\langle\alpha| + \text{H.c.}$, where H.c. indicates Hermitian

[1]Center for Materials Theory, Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, New Jersey 08854-8019, USA. [2]Department of Physics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. [3]Department of Physics, Massachusetts Institute for Technology, Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA.

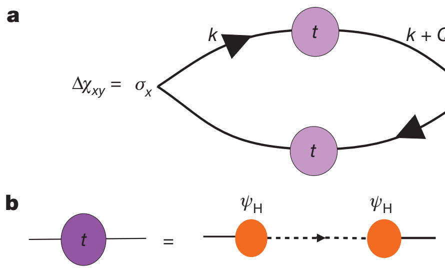


Figure 1 | Phenomenological interpretation of the anomalous spin susceptibility in URu_2Si_2. **a**, The anomalous spin susceptibility is given by conduction electrons (solid lines) scattering off the hidden-order parameter, sandwiched between σ_x and σ_y vertices. **b**, The anomalous scattering *t* matrix can be rewritten as a resonant scattering off the order parameter. The dashed lines represent *f* electrons and the ψ_H vertex represents scattering off the hidden-order parameter.

conjugate, in the low-energy fixed-point Hamiltonian. After two successive time-reversals

$$|k\sigma\rangle V_{\sigma\alpha}(k)\langle\alpha| \rightarrow |k\sigma^{2\pi}\rangle V^{2\pi}_{\sigma\alpha}(k)\langle\alpha^{2\pi}| = -|k\sigma\rangle V^{2\pi}_{\sigma\alpha}(k)\langle\alpha|$$

Because the microscopic Hamiltonian is time-reversal invariant, it follows that $V_{\sigma\alpha}(k) = -V^{2\pi}_{\sigma\alpha}(k)$; the hybridization thus breaks time-reversal symmetry in a fundamentally new way, forming an order parameter that, like a spinor, reverses under 2π rotations. The resulting 'hastatic' order (*hasta* is Latin for spear), is a state of matter that breaks both single and double time-reversal symmetry and is thus distinct from conventional magnetism.

Indirect support for time-reversal symmetry breaking in the hidden-order phase is provided by recent magnetometry measurements that indicate the development of an anisotropic basal-plane spin susceptibility, χ_{xy}, at the hidden-order transition[18]. The strong Ising anisotropy of the *f* electrons prevents them from responding in the basal plane, which leads us to interpret χ_{xy} as a anomalous conduction electron response (Fig. 1), induced by scattering off the hidden-order parameter. In this interpretation, the associated scattering matrix must mix the *x* and *y* components of the conduction electron spins and must take the form $t(k) = (\sigma_x \pm \sigma_y)d(k)$, where $d(k)$ is the scattering amplitude. The scattering matrix $t(k)$ has recently been linked to a spin nematic state[11], under the special condition that $d(-k) = -d^*(k)$ (where an asterisk denotes complex conjugation) to avoid time-reversal symmetry breaking. However, in our interpretation, $d(k)$ is associated with resonant scattering off the Ising *f* state, a process with a real, even-parity scattering amplitude, $d(k) = d(-k)$. In this case, the observed *t* matrix is necessarily odd under time-reversal in the hidden-order phase.

Hybridization in heavy-fermion compounds is usually driven by valence fluctuations mixing a ground-state Kramers doublet and an excited singlet (Fig. 2a). In this case, the hybridization amplitude is a scalar that develops via a crossover, leading to mobile heavy fermions. However, valence fluctuations from a $5f^2$ ground state create excited states with an odd number of electrons and, hence, a Kramers degeneracy (Fig. 2b). Then the quasiparticle hybridization has two components, Ψ_σ, that determine the mixing of the excited Kramers doublet into the ground state. These two amplitudes form a spinor defining the hastatic-order parameter

$$\Psi = \begin{pmatrix} \Psi_\uparrow \\ \Psi_\downarrow \end{pmatrix}$$

The presence of distinct up and down hybridization components indicates that Ψ carries the global spin quantum number; the onset of hybridization must now break double time-reversal and spin rotational invariance by means of a phase transition.

Under pressure, URu_2Si_2 undergoes a first-order phase transition from the hidden-order state to an antiferromagnetic (AFM) state[31]. These two states are remarkably close in energy and share many key features[19,32,33] including common Fermi surface pockets; this motivated the recent proposal that despite the first-order transition separating the two phases, they are linked by 'adiabatic continuity'[32], corresponding to a notional rotation of the hidden order in internal parameter space[5,34]. In the magnetic phase, this spinor points along the *c* axis

$$\Psi_A \propto \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\Psi_B \propto \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

corresponding to time-reversed configurations on alternating layers A and B, implying a large staggered Ising moment. For the hidden-order state, the spinor points in the basal plane

$$\Psi_A \approx \frac{1}{\sqrt{2}}\begin{pmatrix} e^{-i\phi/2} \\ e^{i\phi/2} \end{pmatrix}$$

$$\Psi_B \approx \frac{1}{\sqrt{2}}\begin{pmatrix} -e^{-i\phi/2} \\ e^{i\phi/2} \end{pmatrix}$$

where, again, $\Psi_B = \Theta\Psi_A$, and the hidden order is protected from developing a large moment by the pure Ising character of the $5f^2$ ground state.

Hastatic order permits a direct realization of the adiabatic continuity between the hidden-order and AFM phases in terms of a single Landau functional for the free energy

$$f[T,P,B_z] = [\alpha(T_{HO} - T) - \eta_z B_z^2]|\Psi|^2 + \beta|\Psi|^4 - \gamma(\Psi^\dagger \sigma_z \Psi)^2$$

where $\gamma = \delta(P - P_c)$ (δ being the Dirac delta function) is a pressure-tuned anisotropy term and a dagger denotes adjoint. The unique feature of the theory is that the non-Kramers doublet has Ising

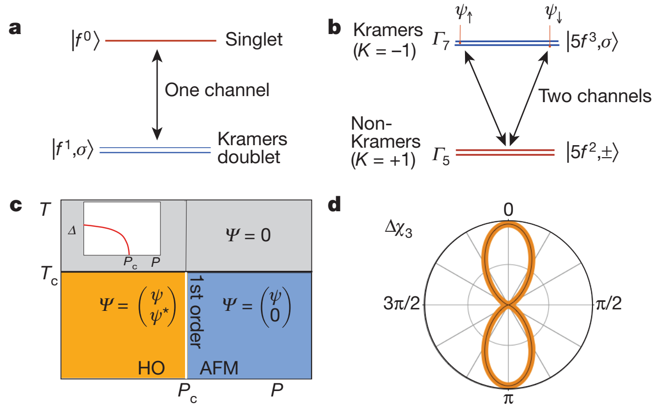


Figure 2 | Spinor hybridization and signatures of hastatic order. **a**, A normal Kondo effect occurs in ions with an odd number of *f* electrons, where the ground state is guaranteed to be doubly degenerate by time-reversal symmetry (and is known as a Kramers doublet). Virtual valence fluctuations to an excited singlet state are associated with a scalar hybridization. **b**, In URu_2Si_2, quasiparticles inherit an Ising symmetry from a $5f^2$ non-Kramers doublet. Loss or gain of an electron necessarily leads to an excited Kramers doublet, and the development of a coherent hybridization is associated with a two-component spinor hybridization that carries a magnetic quantum number and must therefore develop at a phase transition. **c**, Phase diagram for hastatic order, showing how tuning the parameter $\lambda \propto P - P_c$ leads to a spin flop between hastatic order and Ising magnetic order. Inset: at the first-order line, the longitudinal spin gap, Δ, is predicted to vanish because $\Delta \propto \sqrt{P_c - P}$. **d**, Polar plot showing the predicted $\cos^4(\theta)$ form of the nonlinear susceptibility, χ_3, induced by hastatic order, where θ is the angle between the magnetic field and the *c* axis.

character and couples only to the z component of the magnetic field, $B_z = B\cos(\theta)$. The resulting Ising splitting of the non-Kramers doublet suppresses the Kondo effect, giving rise to the B_z^2 term in the quadratic coefficient, where the coefficient η_z is of order $1/T_{\mathrm{HO}}^2$ (Supplementary Information). The phase diagram predicted by this free energy is shown in Fig. 2c. When $P < P_c(T)$, the vector $\Psi^\dagger\vec{\sigma}\Psi = |\Psi|^2(n_x, n_y, 0)$ (where $\vec{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$) denotes a vector of the three Pauli matrices) lies in the basal plane, resulting in hastatic order. At $P = P_c$, there is a first-order 'spin flop' into a magnetic state where $\Psi^\dagger\vec{\sigma}\Psi = |\Psi|^2(0,0,\pm 1)$ lies along the c axis.

Adiabatic continuity provides a natural interpretation of the soft, or low-energy, longitudinal spin fluctuations observed to develop in the hidden-order state[35] as an incipient Goldstone excitation between the two phases[34]. In the hidden-order state, rotations between hastatic and AFM order will lead to a gapped Ising collective mode that we identify with the longitudinal spin fluctuations observed in inelastic neutron scattering[35]. At the first-order phase transition, where $P = P_c$, the quartic anisotropy term vanishes; we predict that the gap, Δ, to longitudinal spin fluctuations will soften according to $\Delta \propto \sqrt{\gamma|\Psi|^2} \approx |\Psi|\sqrt{P_c(T) - P}$ (Supplementary Information). Experimental observation of this feature would provide confirmation of the common origin of the hidden and AFM order.

Another prediction of the phenomenological theory is the development of a nonlinear susceptibility anomaly with a colossal Ising anisotropy. From the Landau theory (Supplementary Information), the jump in the specific heat, ΔC, the susceptibility anomaly, $\mathrm{d}\chi_1/\mathrm{d}T$, and the nonlinear susceptibility anomaly, $\Delta\chi_3$, obey the relation $(\Delta C/T_{\mathrm{HO}})\Delta\chi_3 = 12(\mathrm{d}\chi_1/\mathrm{d}T)^2$, where $\mathrm{d}\chi_1/\mathrm{d}T = -(\alpha/2\beta)\eta_z\cos^2(\theta)$, such that $\Delta\chi_3 \propto \cos^4(\theta)$ (Fig. 2d). A large anomaly in the c-axis nonlinear susceptibility of URu_2Si_2 has been observed at T_{HO} (refs 21, 36), but its Ising anisotropy has never been quantified. The development of a colossal Ising anisotropy in the zero-field nonlinear susceptibility at the hidden-order transition is predicted to be another consequence of hastatic order.

Two-channel valence fluctuation model

We now present a model that relates hastatic order to the valence fluctuations in URu_2Si_2 and is based on a two-channel Anderson lattice model. The uranium ground state is a $5f^2$ Ising Γ_5 doublet[4], $|\pm\rangle = a|\pm 3\rangle + b|\mp 1\rangle$, written in terms of $J = 5/2$ f electrons in the three tetragonal orbitals $\Gamma_7^\pm$ and Γ_6:

$$|+\rangle = \left(af^\dagger_{\Gamma_7^-\downarrow}f^\dagger_{\Gamma_7^+\downarrow} + bf^\dagger_{\Gamma_6\uparrow}f^\dagger_{\Gamma_7^+\uparrow}\right)|0\rangle$$

$$|-\rangle = \left(af^\dagger_{\Gamma_7^-\uparrow}f^\dagger_{\Gamma_7^+\uparrow} + bf^\dagger_{\Gamma_6\downarrow}f^\dagger_{\Gamma_7^+\downarrow}\right)|0\rangle$$

The lowest-lying excited state is most likely the $5f^3$ ($J = 9/2$) state, but for simplicity here we take it to be the symmetry-equivalent $5f^1$ state. Valence fluctuations from the ground state ($5f^2$ Γ_5) to the excited state ($5f^1$ $\Gamma_7^\pm$) occur in two orthogonal conduction channels[37,38], Γ_7^- and Γ_6. This allows us to read off the hybridization matrix elements of the Anderson model

$$H_{\mathrm{VF}}(j) = V_6 c^\dagger_{\Gamma_6\pm}(j)|\Gamma_7^+\pm\rangle\langle\Gamma_5\pm| + V_7 c^\dagger_{\Gamma_7\mp}(j)|\Gamma_7^+\mp\rangle\langle\Gamma_5\pm| + \mathrm{H.c.}$$

where the plus and minus signs respectively denote the 'up' and 'down' states of the coupled Kramers and non-Kramers doublets. The field $c^\dagger_{\Gamma\sigma}(j) = \sum_k \left[\Phi^\dagger_\Gamma(k)\right]_{\sigma\tau} c^\dagger_{k\tau}\mathrm{e}^{-ik\cdot R_j}$ creates a conduction electron at site j (at position R_j) with spin σ in a Wannier orbital with symmetry $\Gamma \in \{\Gamma_6, \Gamma_7\}$, and V_6 and V_7 are the corresponding hybridization strengths. The full model is then written

$$H = \sum_{k\sigma}\epsilon_k c^\dagger_{k\sigma}c_{k\sigma} + \sum_j\left[H_{\mathrm{VF}}(j) + H_a(j)\right]$$

where $H_a(j) = \Delta E\sum_\pm |\Gamma_7^+\pm, j\rangle\langle\Gamma_7^+\pm, j|$ is the atomic Hamiltonian.

Hastatic order is revealed by factorizing the Hubbard operators, $|\Gamma_7^+\sigma\rangle\langle\Gamma_5\alpha| = \hat{\Psi}^\dagger_\sigma\chi_\alpha$. Here $|\Gamma_5\alpha\rangle = \chi^\dagger_\alpha|\Omega\rangle$ is the non-Kramers doublet, represented by the pseudo-fermion $\chi^\dagger_\alpha$, and $\hat{\Psi}^\dagger_\sigma$ is a slave boson[39] representing the excited f^1 doublet $|\Gamma_7^+\sigma\rangle = \hat{\Psi}^\dagger_\sigma|\Omega\rangle$. Hastatic order is the condensation of this boson, $\Psi^\dagger_\sigma\chi_\alpha \to \langle\hat{\Psi}_\sigma\rangle\chi_\alpha$, generating a hybridization between the conduction electrons and the Ising $5f^2$ state while also breaking double time-reversal symmetry. The Γ_5 doublet has both magnetic and quadrupolar moments represented by $\chi^\dagger\vec{\sigma}\chi = (\mathcal{O}_{x^2-y^2}, \mathcal{O}_{xy}, m^z)$, where m^z is the Ising magnetic moment and $\mathcal{O}_{x^2-y^2}$ and $\mathcal{O}_{xy}$ are quadrupole moments. The tensor product $Q_{\alpha\beta} \equiv \Psi_\alpha\Psi^\dagger_\beta$ describes the development of composite order between the non-Kramers doublet and the spin density of conduction electrons. Composite order has been considered previously by several authors in the context of two-channel Kondo lattices[37,40,41] in which the valence fluctuations have been integrated out. However, by factorizing the composite order in terms of the spinor Ψ_α, we are able to understand directly the development of coherent Ising quasiparticles and the broken double time-reversal symmetry.

Using this factorization, we can rewrite the valence fluctuation term as

$$H_{\mathrm{VF}}(j) = \sum_k c^\dagger_{k\sigma}\hat{\mathcal{V}}_{\sigma\eta}(k,j)\chi_\eta(j)\mathrm{e}^{-ik\cdot R_j} + \mathrm{H.c.}$$

where $\hat{\mathcal{V}}(k,j) = V_6\Phi^\dagger_{\Gamma_6}(k)\hat{B}^\dagger_j + V_7\Phi^\dagger_{\Gamma_7^-}(k)\hat{B}^\dagger_j\sigma_1$ with

$$\hat{B}_j = \begin{pmatrix}\hat{\Psi}_\uparrow & 0\\ 0 & \hat{\Psi}_\downarrow\end{pmatrix}$$

In the ordered state, $B_j = \langle\hat{B}_j\rangle$ is replaced by its expectation value, such that in the hidden-order state

$$\left\langle B_j^\dagger\right\rangle = |\Psi|\begin{pmatrix}\mathrm{e}^{i(Q\cdot R_j+\phi)/2} & 0\\ 0 & \mathrm{e}^{-i(Q\cdot R_j+\phi)/2}\end{pmatrix} \equiv |\Psi|U_j$$

with magnitude $|\Psi|$. The internal angle, ϕ, rotates B_j within the basal plane.

Because the hidden-order and AFM phases seem to share a single commensurate wavevector, $Q = (0,0,2\pi/c)$ (refs 19, 32, 33), we use this wavevector here. It is convenient to absorb the unitary matrix U_j into the pseudo-fermion, such that $\tilde{\chi}_j = U_j\chi_j$. This gauge transformation transfers the charge from the slave boson to the pseudo-fermion, making it a charged quasiparticle. In this gauge, one channel (Γ_6) is uniform whereas the other (Γ_7^-) is staggered, and the valence fluctuation Hamiltonian becomes

$$H_{\mathrm{VF}} = \sum_k c^\dagger_k\mathcal{V}_6(k)\chi_k + c^\dagger_k\mathcal{V}_7(k)\chi_{k+Q} + \mathrm{H.c.}$$

where the hybridization form factors are $\mathcal{V}_7(k) = V_7\Phi^\dagger_7(k)\sigma_1$ and $\mathcal{V}_6(k) = V_6\Phi^\dagger_6(k)$

g-factor anisotropy

There are two general aspects of this condensation that deserve special comment. First, the two-channel Anderson impurity model is known to possess a non-Fermi liquid ground state with an entanglement entropy of $(1/2)k_B\ln(2)$ (ref. 42). The development of hastatic order in the lattice liberates this zero-point entropy, accounting naturally for the large entropy of condensation. As a slave boson, Ψ carries both the charge, e, of the electrons and the local gauge charge, $Q_j = \Psi^\dagger_j\Psi_j + \chi^\dagger_j\chi_j$, of constrained valence fluctuations, and its condensation gives a mass to their relative phase through the Higg's

mechanism[43]. But as a Schwinger boson, Ψ's condensation breaks the SU(2) spin symmetry. In this way the hastatic boson can be regarded as a magnetic analogue of the Higg's boson.

One of the key elements of the hastatic theory is the formation of mobile Ising quasiparticles, and the observed Ising anisotropy enables us to set some of the parameters of the theory. The full anisotropic g-factor is a combination of f-electron and conduction-electron components given by $g(\theta)\approx g\cos\theta+g_{\rm c}\left(\frac{T_{\rm K}}{D}\right)$ where $g=2.6$, $g_{\rm c}=2$ and the factor $T_{\rm K}/D$ ($T_{\rm K}$, Kondo temperature; D, conduction-electron bandwidth) derives from the small conduction-electron admixture in the quasiparticles. Experimentally[20], the g-factor anisotropy, that is, the ratio of the c-axis and basal-plane g-factors, $g_z/g_\perp\approx D/T_{\rm K}$, is in excess of 30, which enables us phenomenologically to set a lower bound on $D/T_{\rm K}$ in our model. The g-factor is defined in terms of the Zeeman splitting of the heavy-fermion dispersion, $\Delta E_{k\eta}=|E_{k\eta\uparrow}-E_{k\eta\downarrow}|=g_{k\eta}(\theta)\mu_{\rm B}B$, where $\eta\in[1,4]$ is a band index. Figure 3a shows the Fermi-surface-averaged g-factor, defined by

$$\bar{g}(\theta)=\frac{\sum_{k\eta}g_{k\eta}(\theta)\delta(E_{k\eta})}{\sum_{k\eta}\delta(E_{k\eta})}$$

and calculated within the mean-field hastatic model, as a function of field angle to the c axis, choosing the lower-bound estimate $D/T_{\rm K}\approx 30$.

Broken time-reversal and nematicity

Another key aspect of the hastatic picture is that there must be time-reversal symmetry breaking in both the hidden-order and the AFM phases, manifested by a staggered moment; in the AFM phase this leads to a large c-axis f-electron moment, but in the hidden-order phase it becomes a small transverse moment carried by conduction electrons, $\vec{m}_{\rm c}=-g\mu_{\rm B}{\rm Tr}\vec{\sigma}\mathcal{G}^{\rm c}(k,k+Q)$, where $\mathcal{G}^{\rm c}$ is the conduction-electron Green's function (Supplementary Information). The small magnitude of the induced moment is a consequence of the Clogston–Anderson compensation theorem, which states that changes in the conduction-electron magnetization are set by the same ratio, $T_{\rm K}/D$, that determines the g-factor anisotropy[44]. There will also be a small mixed-valent contribution from the excited Kramers doublet, $\vec{m}_1\propto\langle\Psi^\dagger\vec{\sigma}\Psi\rangle$. The angle of the moments in the plane is controlled by the internal hastatic angle, ϕ. Figure 3b shows the temperature dependence of the in-plane magnetic moment, $m_\perp$, calculated for a case where $D/T_{\rm K}\approx 30$, for which $m_\perp(0)=0.015\mu_{\rm B}$, an upper bound for the predicted conduction-electron moment. Neutron scattering measurements on URu_2Si_2 have placed bounds on the c-axis magnetization of the f electrons using a momentum transfer, Q, in the basal plane. Detection of an $m_\perp(0)$ carried by conduction electrons, with a small scattering form-factor, will require high-resolution measurements with a c-axis momentum transfer. We note that there have been reports from muon spin relaxation and NMR measurements[45,46] of very small intrinsic basal-plane fields in URu_2Si_2, which are consistent with this theory.

Although the conduction electrons develop a magnetic moment, in the hastatic-ordered state, the non-Kramers $5f^2$ state does not develop an ordered dipole or quadrupolar moment, because both the z component and the transverse moment of the pseudovector $\langle\chi^\dagger\vec{\sigma}\chi\rangle=0$ identically vanish. In the microscopic model, the quadrupolar moments vanish because of the d-wave form factor between the Γ_6 and Γ_{7^-} channels (Supplementary Information). The absence of a charge quadrupole implies that there will be no lattice distortion associated with hastatic order. By contrast, hastatic order does induce a weak broken tetragonal symmetry in the spin channel. In the hidden-order state, the interchannel components of the hastatic t matrix, $\hat{\mathcal{V}}_6\hat{\mathcal{V}}_7^\dagger\propto\sigma_x+\sigma_y$, break tetragonal symmetry in the spin channel, resulting in a non-zero spin susceptibility within the conduction fluid

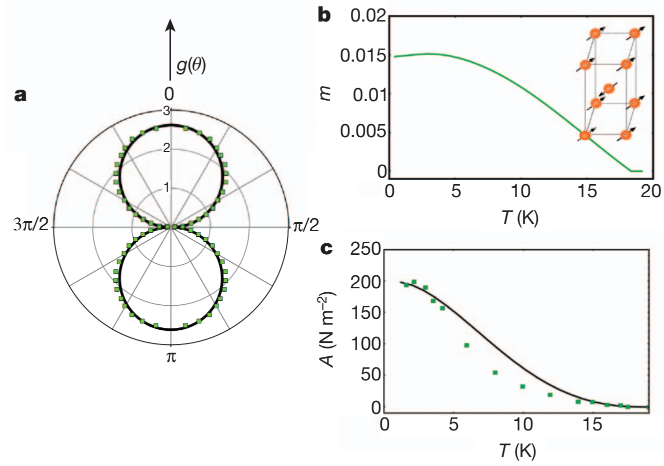


Figure 3 | Magnetic response of hastatic order. a, Polar plot of the calculated g-factor, $g(\theta)$, averaged over the Fermi surface, as a function of magnetic field angle, θ (see Supplementary Information for details), compared with results of ref. 20 (overlaid in green). **b**, As a consequence of the broken time-reversal symmetry, we predict a staggered conduction-electron moment that onsets at the HO transition with a linear, $T_{\rm HO}-T$ temperature dependence (staggering pattern shown in inset). In the plot, the moment is expressed in Bohr magnetons per formula unit. The magnitude of this moment is governed by $T_{\rm K}/D\approx 0.01\mu_{\rm B}/U$, and its orientation is fixed by the way the uniform magnetic susceptibility breaks tetragonal symmetry. **c**, We calculated the tetragonal symmetry breaking component of the uniform susceptibility, $\chi_{xy}(T)$. To compare our results with those of ref. 18 (overlaid in green), we plotted the twofold oscillation amplitude of the magnetic torque, A (black), where $A\cos(2\phi)\equiv\tau_{2\phi}/V=-(\mu_0H^2/V)\cos(2\phi)\chi_{xy}(T)$. This amplitude is proportional to $(T_{\rm HO}-T)^2$ just below the hidden-order transition. For details of our calculation, including parameter choices, see Supplementary Information.

$$\chi_{xy}=-(g\mu_{\rm B})^2{\rm Tr}\sigma_x\mathcal{G}^{\rm c}(k,k+Q)\sigma_y\mathcal{G}^{\rm c}(k+Q,k)$$
$$\propto\left({\rm Tr}\hat{\mathcal{V}}_6\hat{\mathcal{V}}_7^\dagger\right)^2$$

of a magnitude of order $(T_{\rm K}/D)^2$, which onsets at the hidden-order temperature as $|\Psi|^4\approx(T_{\rm HO}-T)^2$ as shown in Fig. 3c.

Hastatic order also manifests itself as an anisotropic hybridization gap, which vanishes along lines in momentum space, giving rise to a V-shaped density of states due to the partial gapping of the Fermi surface, as shown in Fig. 4, which will be smeared out by disorder in the real material. The anisotropy of the hybridization also breaks tetragonal symmetry, giving rise to a energy-dependent nematicity, $\eta(E)$, that peaks over a narrow energy window around the Kondo resonance. Some of this nematicity is present at the Fermi surface, accounting for the splitting seen in quantum oscillation frequencies[22] and cyclotron resonance experiments[47]. An ideal way to verify this prediction is to use scanning tunnelling spectroscopy, where the measured differential conductivity, ${\rm d}I/{\rm d}V\propto A(eV,x)$ is proportional to the local density of states $A(eV,x)$ at position x on the surface. A measure of the broken tetragonal symmetry is provided by the 'nematicity'

$$\eta(V)=\frac{\overline{({\rm d}I/{\rm d}V)(x,y)\ {\rm sgn}(xy)}}{\left(\overline{({\rm d}I/{\rm d}V)^2}-\left(\overline{{\rm d}I/{\rm d}V}\right)^2\right)^{1/2}}$$

Here x and y are the coordinates relative to the centre of the unit cell and the overbar denotes an average over the unit cell. The resonant scattering of the hastatic order causes this quantity to vary rapidly as a function of voltage, over an energy range of order the Kondo temperature $T_{\rm K}$. Figure 4 shows the variation of the nematicity, calculated within our model of hastatic order. The nematicity is found to peak at the Kondo resonance, at a value of approximately 50%.

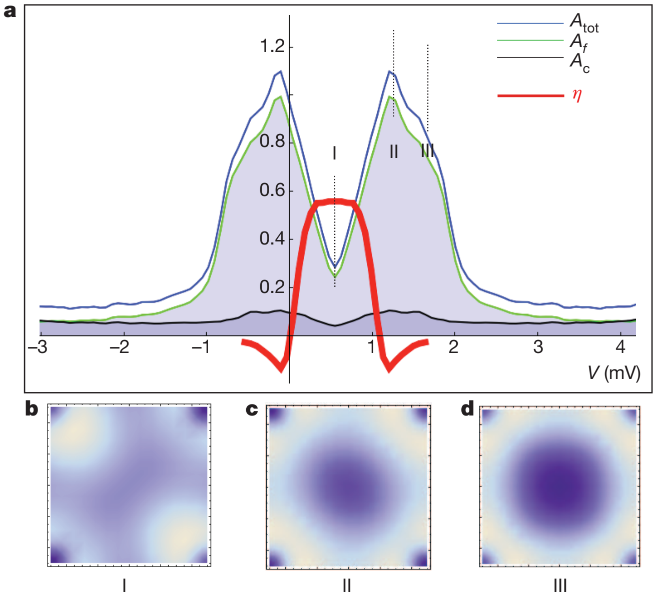


Figure 4 | Density of states and resonant nematicity predicted by theory. **a**, Density of states (arbitrary units) as a function of energy predicted by model calculation (blue line), showing f-electron and conduction-electron components. Red line, voltage dependence of nematicity, $\eta(V)$, in model calculation of scanning tunnelling spectrum. **b**–**d**, Spatial dependence of density of states for selected bias voltages in model calculation of scanning tunnelling spectrum, showing the resonant character of the nematicity. Voltages I, II and III are located at the center, the maximum and the shoulder of the density of states, respectively.

Beyond mean-field theory

So far we have focused on the mean-field consequences of hastatic order. The recently observed softening of the commensurate longitudinal spin fluctuations at T_{HO} (ref. 48) suggests that hastatic order 'melts' via phase fluctuations (the hastatic-order parameter vanishes but correlations remain). Although the hybridization spinor itself, $\langle\Psi\rangle$, can become non-zero only below the phase transition at T_{HO}, we expect that its amplitude, $\langle\Psi^\dagger\Psi\rangle$, will have a non-zero value to higher temperatures. Because $\langle\Psi^\dagger\vec{\sigma}\Psi\rangle$ remains zero above T_{HO}, only the non-symmetry-breaking (uniform, intrachannel) components of the hybridization can develop: $\hat{\mathcal{V}}_6\hat{\mathcal{V}}_6^\dagger$ and $\hat{\mathcal{V}}_7\hat{\mathcal{V}}_7^\dagger$ will emerge via a crossover at a higher temperature, T^*, to create an incoherent Fermi liquid, consistent with the heavy mass inferred from thermodynamic and optical measurements[1,17] and the development of Fano signatures in both scanning and point-contact tunnelling spectroscopy[14–16]. The symmetry-breaking, interchannel components, $\hat{\mathcal{V}}_6\sigma_1\hat{\mathcal{V}}_7^\dagger$, will always develop precisely at the hidden-order transition. Another aspect of experiments that is not covered by our mean-field description is the observation of gapped, low-energy incommensurate fluctuations around a Q-vector $Q^* = (1 \pm 0.4, 0, 0)$ in the hidden-order phase[33,35,48,49], which seems to be a sign of an unfulfilled predisposition towards an incommensurate phase, probably driven by partial Fermi surface nesting. These effects lie beyond a mean-field description, but would emerge from the Gaussian fluctuations about the mean-field theory.

Although we have discussed hastatic order in the context of URu_2Si_2, it should be a more widespread phenomenon associated with hybridization in any f-electron material whose unfilled f shell contains a geometrically stabilized non-Kramers doublet. As such, we expect realizations of hastatic order in other $5f$ uranium and $4f$ praseodymium intermetallic compounds.

Received 17 July; accepted 27 November 2012.

1. Palstra, T. T. M. *et al.* Superconducting and magnetic transitions in the heavy-fermion system URu_2Si_2. *Phys. Rev. Lett.* **55,** 2727–2730 (1985).
2. Schlabitz, W. *et al.* Superconductivity and magnetic order in a strongly interacting Fermi system: URu_2Si_2. *Z. Phys. B* **62,** 171–177 (1986).
3. Mydosh, J. A. & Oppeneer, P. M. Hidden order, superconductivity and magnetism – the unsolved case of URu_2Si_2. *Rev. Mod. Phys.* **83,** 1301–1322 (2011).
4. Amitsuka, H. & Sakakibara, T. Single uranium-site properties of the dilute heavy electron system $U_xTh_{1-x}Ru_2Si_2$ ($x \leq 0.07$). *J. Phys. Soc. Jpn* **63,** 736–747 (1994).
5. Haule, K. & Kotliar, G. Arrested Kondo effect and hidden order in URu_2Si_2. *Nature Phys.* **5,** 796–799 (2009).
6. Santini, P. & Amoretti, G. Crystal field model of the magnetic properties of URu_2Si_2. *Phys. Rev. Lett.* **73,** 1027–1030 (1994).
7. Varma, C. M. & Zhu, L. Helicity order: hidden order parameter in URu_2Si_2. *Phys. Rev. Lett.* **96,** 036405–036408 (2006).
8. Pépin, C., Norman, M. R., Burdin, S. & Ferraz, A. Modulated spin liquid: a new paradigm for URu_2Si_2. *Phys. Rev. Lett.* **106,** 106601–106604 (2011).
9. Yuan, T., Figgins, J. & Morr, D. K. Hidden order transition in URu_2Si_2: evidence for the emergence of a coherent Anderson lattice from scanning tunneling spectroscopy. *Phys. Rev. B* **86,** 035129–035134 (2012).
10. Dubi, Y. & Balatsky, A. V. Hybridization wave as the 'hidden order' in URu_2Si_2. *Phys. Rev. Lett.* **106,** 086401–086404 (2011).
11. Fujimoto, S. Spin nematic state as a candidate of the hidden order phase of URu_2Si_2. *Phys. Rev. Lett.* **106,** 196407–196410 (2011).
12. Ikeda, H. *et al.* Emergent rank-5 'nematic' order in *URusi₂*. *Nature Phys.* **8,** 528–533 (2012).
13. Santander-Syro, A. F. *et al.* Fermi-surface instability at the 'hidden order' transition of URu_2Si_2. *Nature Phys.* **5,** 637–641 (2009).
14. Schmidt, A. R. *et al.* Imaging the Fano lattice to 'hidden order' transition in URu_2Si_2. *Nature* **465,** 570–576 (2010).
15. Aynajian, P. *et al.* Visualizing the formation of the Kondo lattice and the hidden order in URu_2Si_2. *Proc. Natl Acad. Sci. USA* **107,** 10383–10388 (2010).
16. Park, W. K. *et al.* Fano resonance and hybridization gap in Kondo lattice URu_2Si_2. *Phys. Rev. Lett.* **108,** 246403 (2012).
17. Nagel, U. *et al.* Optical spectroscopy shows that the normal state of URu_2Si_2 is an anomalous Fermi liquid. *Proc. Natl Acad. Sci. USA* **109,** 19161–19165 (2012).
18. Okazaki, R. *et al.* Rotational symmetry breaking in the hidden order phase of *URu_2Si_2*. *Science* **331,** 439–442 (2011).
19. Hassinger, E. *et al.* Similarity of the Fermi surface in the hidden order state and in the antiferromagnetic state of URu_2Si_2. *Phys. Rev. Lett.* **105,** 216409–216412 (2010).
20. Altarawneh, M. M. *et al.* Sequential spin polarization of the Fermi surface pockets of URu_2Si_2 and its implications for the hidden order. *Phys. Rev. Lett.* **106,** 146403–146416 (2011).
21. Ramirez, A. P. *et al.* Nonlinear susceptibility as a probe of tensor spin order in URu_2Si_2. *Phys. Rev. Lett.* **68,** 2680–2683 (1992).
22. Ohkuni, H. *et al.* Fermi surface properties and de Haas-van Alphen oscillation in both the normal and superconducting mixed states of URu_2Si_2. *Philos. Mag. B* **79,** 1045–1077 (1999).
23. Brison, J. P. *et al.* Anisotropy of the upper critical field in URu_2Si_2 and FFLO state in antiferromagnetic superconductors. *Physica C* **250,** 128–138 (1995).
24. Altarawneh, M. M. *et al.* Superconducting pairs with extreme uniaxial anisotropy in URu_2Si_2. *Phys. Rev. Lett.* **108,** 066407–066410 (2012).
25. Goremychkin, E. A. *et al.* Magnetic correlations and the anisotropic Kondo effect in $Ce_{1-x}La_xAl_3$. *Phys. Rev. Lett.* **89,** 147201–147204 (2002).
26. Flint, R., Chandra, P. & Coleman, P. Basal-plane nonlinear susceptibility: a direct probe of the single-ion physics in URu_2Si_2. *Phys. Rev. B* **86,** 155155–155160 (2012).
27. Nieuwenhuys, G. J. Crystalline electric field effects in UPt_2Si_2 and URu_2Si_2. *Phys. Rev. B* **35,** 5260–5263 (1987).
28. Zołnierek, Z. & Troc, R. Magnetic properties of tetragonal uranium componds. I. The U_2N_2Z ternaries. *J. Magn. Magn. Mater.* **8,** 210–222 (1978).
29. Ohkawa, F. J. & Shimizu, H. Quadrupole and dipole orders in URu_2Si_2. *J. Phys. Condens. Matter* **11,** L519–L524 (1999).
30. Sakurai, J. J. *Modern Quantum Mechanics* rev. edn 266–282 (Addison-Wesley, 1994).
31. Amitsuka, H. *et al.* Pressure-temperature phase diagram of the heavy-electron superconductor URu_2Si_2. *J. Magn. Magn. Mater.* **310,** 214–220 (2007).
32. Jo, Y. J. *et al.* Field-induced Fermi surface reconstruction and adiabatic continuity between antiferromagnetism and the hidden-order state in URu_2Si_2. *Phys. Rev. Lett.* **98,** 166404 (2007).
33. Villaume, A. *et al.* Signature of hidden order in heavy fermion superconductor URu_2Si_2: resonance at the wave vector $Q_0 = (1, 0, 0)$. *Phys. Rev. B* **78,** 012504 (2008).
34. Haule, K. & Kotliar, G. Complex Landau-Ginzburg theory of the hidden order in URu_2Si_2. *Europhys. Lett.* **89,** 57006 (2010).
35. Broholm, C. *et al.* Magnetic excitations in the heavy-fermion superconductor URu_2Si_2. *Phys. Rev. B* **43,** 12809–12822 (1991).
36. Miyako, Y. *et al.* Magnetic properties of $U(Ru_{1-x}Rh_x)_2Si_2$ single crystals ($0 \leq x \leq 1$). *J. Appl. Phys.* **70,** 5791 (1991).
37. Cox, D. L. & Jarrell, M. The two-channel Kondo route to non-Fermi liquids. *J. Phys. Condens. Matter* **8,** 9825–9853 (1996).
38. Cox, D. L. & Zawadowski, A. *Exotic Kondo Effects in Metals* (Taylor & Francis, 2002).
39. Coleman, P. A new approach to the mixed valence problem. *Phys. Rev. B* **29,** 3035–3044 (1984).
40. Coleman, P., Tsvelik, A. M., Andrei, N. & Kee, H. Y. Co-operative Kondo effect in the two-channel Kondo lattice. *Phys. Rev. B* **60,** 3608–3628 (1999).
41. Hoshino, S., Otsuki, J. & Kuramoto, Y. Diagonal composite order in a two-channel Kondo lattice. *Phys. Rev. Lett.* **107,** 247202–247205 (2011).

42. Bolech, C. & Andrei, N. Solution of the two-channel Anderson impurity model: implications for the heavy fermion UBe_{13}. *Phys. Rev. Lett.* **88,** 237206–237209 (2002).
43. Coleman, P., Marston, J. B. & Schofield, A. J. Transport anomalies in a simplified model for a heavy-electron quantum critical point. *Phys. Rev. B* **72,** 245111 (2003).
44. Anderson, P. W. Localized magnetic states in metals. *Phys. Rev.* **124,** 41–53 (1961).
45. Amitsuka, H. *et al.* Inhomogeneous magnetism in URu_2Si_2 studied by muon spin relaxation under high pressure. *Physica B* **326,** 418–421 (2003).
46. Bernal, O. O. *et al.* Ambient pressure ^{99}Ru NMR in URu_2Si_2: internal field anisotropy. *J. Magn. Magn. Mater.* **272–276,** E59–E60 (2004).
47. Tonegawa, S. *et al.* Cyclotron resonance in the hidden-order phase of URu_2Si_2. *Phys. Rev. Lett.* **109,** 036501 (2012).
48. Niklowitz, P. G. *et al.* Role of commensurate and incommensurate low-energy excitations in the paramagnetic to hidden-order transition of URu_2Si_2. Preprint at http://arxiv.org/abs/1110.5599 (2011).
49. Wiebe, C. R. *et al.* Gapped itinerant spin excitations account for missing entropy in the hidden order state of URu_2Si_2. *Nature Phys.* **3,** 96–99 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements An early version of this work was begun in collaboration with P. Fazekas, since deceased. We thank N. Andrei, S. Burdin, B. Coleman, L. Greene, N. Harrison, K. Haule, G. Kotliar, P. Lee, G. Luke, Y. Matsuda, J. Mydosh, P. Niklowitz, C. Pépin, T. Senthil, A. Toth and T. Timusk for discussions. We acknowledge funding from the Simons Foundation (R.F.), the US National Science Foundation grant DMR 0907179 (R.F., P. Coleman), the US National Science Foundation I2CAM International Materials Institute Award Grant DMR-0844115 (R.F., P. Coleman) and the US National Science Foundation grant 1066293 (all authors) while at the Aspen Center for Physics. We are grateful for the hospitality of the Aspen Center for Physics.

Author Contributions All authors contributed equally in the discussions and development of the hastatic-order concept, the experimental consequences and its mean-field description pertinent to URu_2Si_2. R.F. and P. Coleman carried out the detailed numerical calculations of the microscopic model. All authors contributed towards the writing of the paper and Supplementary Information.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P. Coleman. (coleman@physics.rutgers.edu).

doi:10.1038/nature11647

Towards germline gene therapy of inherited mitochondrial diseases

Masahito Tachibana[1], Paula Amato[2], Michelle Sparman[1], Joy Woodward[1], Dario Melguizo Sanchis[1], Hong Ma[1], Nuria Marti Gutierrez[1], Rebecca Tippner-Hedges[1], Eunju Kang[1], Hyo-Sang Lee[1], Cathy Ramsey[1], Keith Masterson[2], David Battaglia[2], David Lee[2], Diana Wu[2], Jeffrey Jensen[1,3], Phillip Patton[2], Sumita Gokhale[4], Richard Stouffer[1,2] & Shoukhrat Mitalipov[1,2]

Mutations in mitochondrial DNA (mtDNA) are associated with severe human diseases and are maternally inherited through the egg's cytoplasm. Here we investigated the feasibility of mtDNA replacement in human oocytes by spindle transfer (ST; also called spindle–chromosomal complex transfer). Of 106 human oocytes donated for research, 65 were subjected to reciprocal ST and 33 served as controls. Fertilization rate in ST oocytes (73%) was similar to controls (75%); however, a significant portion of ST zygotes (52%) showed abnormal fertilization as determined by an irregular number of pronuclei. Among normally fertilized ST zygotes, blastocyst development (62%) and embryonic stem cell isolation (38%) rates were comparable to controls. All embryonic stem cell lines derived from ST zygotes had normal euploid karyotypes and contained exclusively donor mtDNA. The mtDNA can be efficiently replaced in human oocytes. Although some ST oocytes displayed abnormal fertilization, remaining embryos were capable of developing to blastocysts and producing embryonic stem cells similar to controls.

Mitochondrial DNA is localized in the cell's cytoplasm, whereas chromosomal genes are confined to the nucleus. Each cell may have thousands of mtDNA copies, which may all be mutated (homoplasmy) or exist as a mixture (heteroplasmy). The clinical manifestations of mtDNA diseases vary, but often affect organs and tissues with the highest energy requirements, including the brain, heart, muscle, pancreas and kidney[1]. The expression and severity of disease symptoms depends on the specific mutation and heteroplasmy levels[1].

An estimated prevalence of inherited mtDNA diseases is 1 in every 5,000–10,000 live births, suggesting that, in the United States alone, between 1,000 and 4,000 children are born every year with mtDNA diseases[2,3]. Based on other estimates, the frequency of pathogenic mtDNA mutations is even higher—1 in 200 children inherit mutations[4]. However, not all of these children develop the disease at birth, because mtDNA mutations are present at low heteroplasmy levels.

At present, there are no cures for mitochondrial disorders and available treatments only alleviate symptoms and delay disease progression. Therefore, several strategies for preventing transmission of mtDNA mutations from mothers to their children have been actively pursued.

One approach is to completely replace the mutated mtDNA of a patient's oocyte with the healthy mitochondrial genome from an oocyte donated by another woman using spindle transfer (ST)[5]. The technique isolates and transplants the chromosomes (nuclear genetic material) from a patient's unfertilized oocyte into the cytoplasm of another enucleated egg, containing healthy mtDNA as well as other organelles, RNA and proteins. A child born as a result of the ST procedure will be the genetic child of the patient but carry healthy mitochondrial genes from the donor. Our prior studies in a monkey model demonstrated not only the feasibility of the ST procedure but also that ST is highly effective and compatible with normal fertilization and birth of healthy offspring[6]. This strategy has been considered clinically to be a highly important future gene therapy to avoid transmission of serious mitochondrial diseases (http://www.hfea.gov.uk/6372.html).

Here we present a comprehensive study demonstrating the feasibility and outcomes of ST with human oocytes donated by healthy volunteers. To measure success, we fertilized reconstructed oocytes *in vitro* and assessed the normality of fertilization and embryo development to blastocysts. In addition, we derived embryonic stem cells (ESCs) and carried out detailed genetic analyses to assess efficacy of gene replacement and possible chromosomal abnormalities associated with ST. We also conducted additional studies in a rhesus macaque model to investigate the feasibility of using cryopreserved oocytes for ST and postnatal development of ST offspring.

Mitochondrial DNA replacement in human oocytes

Seven volunteers (aged 21–32 years) underwent ovarian stimulation and a total of 106 mature metaphase II (MII) oocytes were retrieved (range of 7–28, or a mean of 15 oocytes per donor cycle). Participants were synchronized in three separate experiments, so that at least two fresh oocyte cohorts were available on the same day for reciprocal ST. We selected a total of 65 MII oocytes for the ST procedure and 33 served as non-manipulated controls (Fig. 1a). We successfully transferred the spindle surrounded by a membrane and a small amount of cytoplasm (karyoplast) between 64 oocytes (98%; Fig. 1a, b). Sixty oocytes survived fertilization by intracytoplasmic sperm injection (ICSI; 94%) and 44 formed visible pronuclei (73%). These outcomes were similar to the results for controls; 32 oocytes survived ICSI (97%) and 24 (75%) formed pronuclei (Fig. 1b). Microscopic evaluations determined that almost half of ST zygotes (21/44, 48%) contained normal two pronuclei and two polar bodies (2PN/2PB) (Fig. 2a). However, the remaining ST zygotes had an irregular number of pronuclei and/or polar bodies (Fig. 2a). Abnormal fertilization was also observed in the intact control group, albeit at a lower incidence (3/24, 13%).

[1]Division of Reproductive & Developmental Sciences, Oregon National Primate Research Center, Oregon Health & Science University, 505 NW 185th Avenue, Beaverton, Oregon 97006, USA. [2]Division of Reproductive Endocrinology, Department of Obstetrics and Gynecology, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. [3]Center for Women's Health, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. [4]University Pathologists, LLC, Boston University School of Medicine; Roger Williams Medical Center, Providence, Rhode Island 02908, USA.

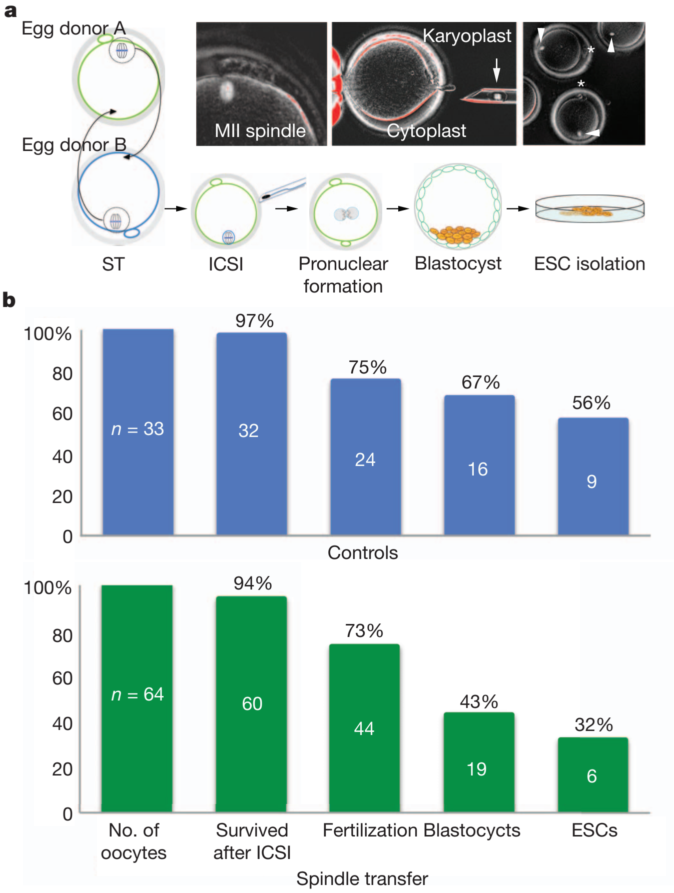


Figure 1 | Experimental design and main outcomes after ST with human oocytes. a, Oocytes were retrieved from two unrelated donors and spindle–chromosomal complexes were reciprocally exchanged. Reconstructed oocytes were fertilized by ICSI and monitored for *in vitro* development to blastocysts and ESCs. Images in boxes depict a human mature MII oocyte with the spindle visualized under polarized microscope (left), isolated cytoplast and karyoplast (middle) and intact spindles inside recipient cytoplasts after transfers (right, arrowheads). Original magnifications: left and middle, ×200; right, ×100. Asterisks indicate first polar bodies. **b**, Experimental outcomes after ST in human oocytes. The top and bottom graphs represent fertilization, blastocyst and ESC isolation rates for intact control and ST embryos, respectively. No statistical differences were found between ST and controls in survival after ICSI, fertilization, blastocyst development and ESC derivation rates ($P > 0.05$).

Blastocyst formation rate in the normally fertilized ST group (13/21, 62%) was statistically similar to controls (16/21, 76%). However, the majority of abnormal ST zygotes arrested, with only 26% (6/23) reaching blastocysts (Supplementary Table 1). Interestingly, blastocyst development of 3PN/1PB zygotes was noticeably higher (4/11, 36%) than that of other abnormally fertilized ST groups (Supplementary Table 1).

Derivation and genetic analysis of human ESCs

To provide additional insights into the developmental competence of ST-produced human embryos and to obtain sufficient material for molecular and cytogenetic analyses, we derived ESCs from blastocysts. Nine ESC lines (HESO lines) were established from 16 control blastocysts (56%; Fig. 1b). This ESC derivation rate is significantly higher than currently reported for embryos donated by IVF patients[7]. Similarly, 13 ST blastocysts developed from normally fertilized zygotes produced five ESC lines (38%; HESO-ST lines). We also plated four ST blastocysts that originated from 3PN/1PB zygotes and derived one ESC line (25%; Supplementary Table 1).

The ESCs derived from the ST embryos had normal morphology and were indistinguishable from controls (Supplementary Fig. 1a, b).

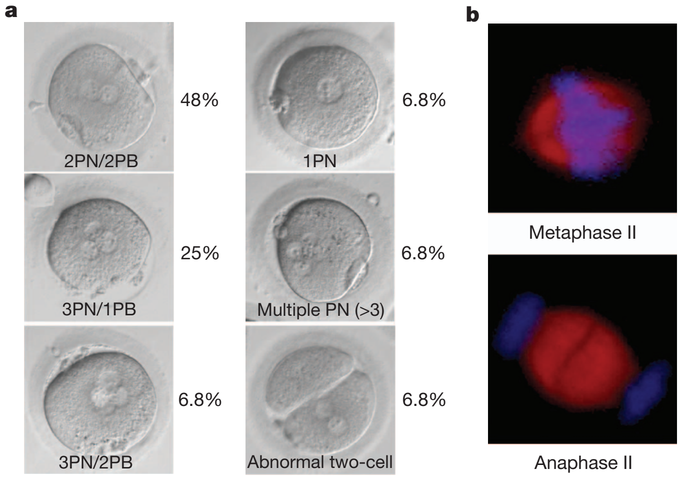


Figure 2 | Abnormal pronuclear formation and spindle morphology in human ST zygotes. a, Proportion of normally fertilized zygotes with two pronuclei and two polar bodies (2PN/2PB) compared with abnormal zygotes (3PN/1PB, 3PN/2PB, 1PN, multiple PN and two-cell) after ST. **b**, Integrity of meiotic spindles in human ST oocytes depicting normal metaphase II (top), and premature progression to the anaphase II (bottom).

All ESCs expressed standard pluripotency markers, including OCT-4, SOX2, SSEA-4, TRA-1-81 and TRA-1-60 (Supplementary Fig. 1a, b). Following injection into immunodeficient mice, experimental ESCs formed teratoma tumours consisting of cells and tissues representing all three germ layers (Supplementary Fig. 1c). Detailed analysis of nuclear DNA using microsatellite markers confirmed that all HESO-ST lines inherited their chromosomes from the spindle donor oocytes (Fig. 3a, Table 1 and Supplementary Fig. 2a, d). Analysis of mtDNA confirmed that all HESO-ST cell lines derived their mtDNA from the cytoplast donors (Fig. 3b, Table 1 and Supplementary Figs 2b, e and 3)[6,8].

As reported previously[6], small amounts of mtDNA are usually co-transferred with the karyoplast during the ST, causing a low carryover heteroplasmy. In clinical situations this may result in the transmission of mutant mtDNA to ST embryos and children. Therefore we conducted both qualitative and quantitative mtDNA assays to determine the degree of mtDNA carryover in ST embryos and ESC lines. We identified mtDNA sequence differences between oocyte donors and unique restriction enzyme recognition sites for restriction-fragment length polymorphism (RFLP) assay[6]. For example, mtDNA of the egg donor 1 (spindle contributor for HESO-ST2, -ST3 and -ST4) possessed a unique *EcoRV* digestion sequence GATATC (Supplementary Fig. 4a). In contrast, egg donor 2 (mtDNA contributor) carried a single nucleotide polymorphism GATACC precluding enzyme recognition. The results confirmed that mtDNA in ST cell lines was exclusively derived from the cytoplast donors with no detectable mtDNA carryover (Table 1 and Supplementary Figs 2c and 4a). We also used more sensitive ARMS–qPCR (amplification refractory mutation system–quantitative polymerase chain reaction) that enables the measurement of heteroplasmy below 1%[9,10]. The mean mtDNA carryover in ST oocytes and embryos was 0.5% (s.d. ± 0.4; range 0–0.9%) and in HESO-ST cell lines was 0.6% (s.d. ± 0.9; range 0–1.7%) (Supplementary Fig. 4b and Supplementary Table 2). These results are consistent with our previous data from a nonhuman primate[6] and suggest negligible mtDNA carryover in ST offspring.

Cytogenetic analyses, using G-banding, indicated that all five HESO-ST lines derived from normal ST embryos contained diploid male or female karyotypes, with no evidence of detectable numerical or structural chromosomal abnormalities (Table 1 and Supplementary Fig. 5). However, two out of nine lines derived from control embryos had numerical aberrations. Notably, HESO-6 carried a 47 XYY karyotype, whereas HESO-9 was 45 XO (Supplementary Table 3 and

a Nuclear DNA genotyping

	AME	D7S513	D6S1691
Sperm donor	XY	191/193	227/229
Egg donor 1	XX	191/199	211/223
Egg donor 2	XX	187/201	225/225
HESO-ST2	XX	193/199	223/229
HESO-ST3	XX	191/193	211/227
HESO-ST4	XY	193/199	211/229

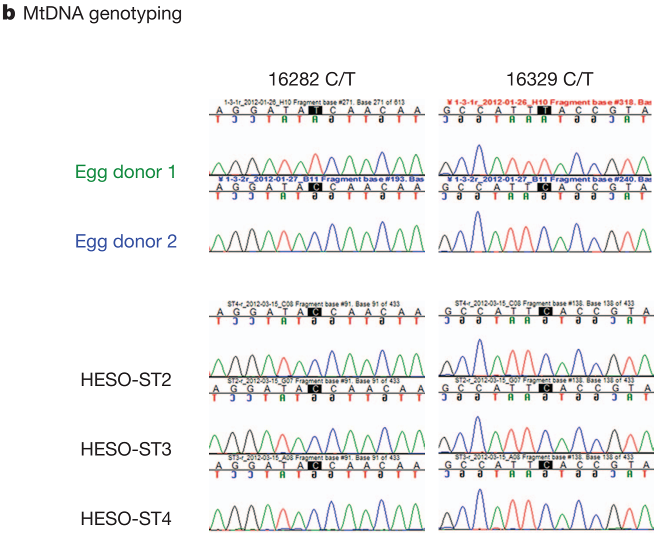


Figure 3 | Genetic analysis of ESCs derived from human ST embryos. **a**, Nuclear DNA origin of HESO-ST2, -ST3 and -ST4 determined by microsatellite parentage analysis. The microsatellite markers for D7S513 and D6S1691 loci demonstrate that the nuclear DNA in these ESC lines was from the egg donor 1 (the spindle donor). **b**, mtDNA genotyping by direct sequencing show that the mtDNA in HESO-ST2, -ST3 and -ST4 is originated from the egg donor 2.

Supplementary Fig. 6). G-banding revealed that HESO-ST6, the cell line derived from abnormally fertilized ST zygote, contained abnormal triploid female chromosome complement (Supplementary Fig. 5). In addition, detailed microsatellite analysis of nuclear DNA in this cell line confirmed the presence of three alleles for most short tandem repeat (STR) loci (Supplementary Table 4). On the basis of allele inheritance, we concluded that the triploid karyotype was caused by retention of the genetic material of the second polar body. This was consistent with observation of the extra pronucleus but lack of the second polar body in the zygote.

Abnormal fertilization in human zygotes produced by ST

Abnormal fertilization observed in some human ST oocytes was unexpected because this was not observed in monkey studies[6]. Therefore, additional experiments were conducted to investigate possible underlying mechanisms. Genetic analysis of the HESO-ST6 cell line hinted that some ST oocytes and embryos retain extra chromosomes that are normally extruded into the second polar body. This was likely to be caused by sub-optimal conditions during the ST procedure that disturbed spindle integrity. Initially, we focused on the effect of cytochalasin B (CB), a microfilament inhibitor known to inhibit cytokinesis and to block the second PB extrusion in oocytes[11,12]. CB is used acutely during the ST procedure and oocytes are thoroughly rinsed before fertilization, but residual CB could interrupt the second PB extrusion. Therefore, we extended incubation time between ST and ICSI, or decreased CB concentration. However, abnormal fertilization persisted even in the absence of CB (Supplementary Tables 5 and 6), indicating that abnormal meiotic segregation is not likely to be caused by CB exposure.

We next addressed whether oocyte polarity during displacement of spindles leads to abnormal meiosis. Typically, spindles in MII oocytes are adjacent to the first polar bodies. However, during the ST procedure, karyoplasts are reintroduced on the opposite side (referred to as 180 degree)[5,6]. We reintroduced spindles next to or at the 90 degree from the first PB (Supplementary Fig. 7). However, ST zygotes had similar pronuclear abnormalities (Supplementary Fig. 7).

Lastly, we reasoned that human meiotic spindles may undergo premature activation during the ST manipulations leading to incomplete resumption of meiosis after fertilization. Spindle morphology and meiotic stage was analysed in intact and ST human oocytes following immunolabelling with α- and β-tubulin. Analysis demonstrated that in some ST oocytes spindles already progressed to the late anaphase II, whereas all control oocytes maintained uniform metaphase II (Fig. 2b).

Oocyte cryopreservation before ST

Current ST protocols use fresh oocytes and require that both patient and healthy mtDNA egg donors undergo synchronous retrievals. However, it is difficult to manage the same-day egg retrievals owing to differences in the ovarian cycle and responses to gonadotropins. In addition, an equal number of patient and donor eggs retrieved would be ideal to avoid oocyte wastage. Therefore, oocyte freezing, storage and thawing will be critical for clinical applications of the ST. Recent advances in oocyte vitrification procedures suggest that cryopreserved human MII oocytes can be used in clinical IVF practice with the same efficiency as fresh eggs[13,14]. To evaluate the feasibility of using cryopreserved oocytes for ST, we turned to the nonhuman primate model. We tested a commercially available vitrification kit (CRYOTOP) and determined that survival and recovery of rhesus macaque MII oocytes post-thaw is high. After ISCI, 72% formed pronuclei, but only 6% developed to blastocysts ($P < 0.05$) (Table 2). Thus, this cryopreservation method compromises blastocyst development, because blastocyst formation of fresh oocytes from the same cohort was 52% (Table 2)[6,9].

We next conducted reciprocal ST between fresh and frozen-thawed monkey oocytes and examined fertilization and embryo development (Supplementary Fig. 8). When fresh spindles were transplanted into vitrified cytoplasts, fertilization after ICSI was impaired (50%)

Table 1 | Genetic analysis of human ESCs derived from ST blastocysts

Cell line	HESO-ST-2	HESO-ST-3	HESO-ST-4	HESO-ST-5	HESO-ST-6	HESO-ST-7
Nuclear donor	Donor 1	Donor 1	Donor 1	Donor 4	Donor 3	Donor 6
Cytoplast donor	Donor 2	Donor 2	Donor 2	Donor 3	Donor 4	Donor 7
Fertilization	2PN/2PB	2PN/2PB	2PN/2PB	2PN/2PB	3PN/1PB	2PN/2PB
Karyotype (passage no.)	46 XX, P4	46 XX, P7	46 XY, P7	46 XX, P4	69 XXX, P4	46 XY, P3
Nuclear DNA origin (by STR)	Donor 1	Donor 1	Donor 1	Donor 4	Donor 3	Donor 6
mtDNA origin	Donor 2	Donor 2	Donor 2	Donor 3	Donor 4	Donor 7
mtDNA carryover (RFLP)	Undetectable	Undetectable	Undetectable	Undetectable	Undetectable	NT
mtDNA carryover (ARMS–qPCR)	0.20%	0.01%	1.70%	NT	NT	NT

NT, not tested.

Table 2 | Fertilization and embryo development of frozen rhesus oocytes

Experiment	Group	*n*	Survived after ST (%)	Survived after ICSI (%)	Fertilized (%)	Blastocysts (%)
1	Fresh oocytes	32	NA	30 (94)	29 (97)	15 (52)*
	Vitrified oocytes	26	NA	25 (96)	18 (72)	1 (6)
2	Control fresh oocytes	34	NA	33 (97)	30 (91)†	17 (57)‡
	Fresh cytoplasts Vitrified spindles	36	34 (94)	32 (94)	28 (88)†	19 (68)‡
	Vitrified cytoplasts Fresh spindles	35	35 (100)	34 (97)	17 (50)	0

* Blastocyst rate statistically different from that for vitrified oocytes ($P < 0.05$).
† Fertilized rate statistically different from that for vitrified cytoplasts with fresh spindles ($P < 0.05$).
‡ Blastocyst rate statistically different from that for vitrified cytoplasts with fresh spindles ($P < 0.05$).
Data were analysed using analysis of variance (ANOVA).
NA, not applicable.

compared to controls (91%) (Table 2). Moreover, all embryos in this ST group arrested before reaching blastocysts, whereas 57% controls progressed to blastocysts. These ST results were similar to those seen with frozen-thawed intact controls (Table 2). However, when spindles from vitrified oocytes were transferred into fresh cytoplasts, fertilization (88%) and blastocyst formation (68%) rates were similar to fresh controls (Table 2). These results indicate that vitrification causes damage primarily within the cytoplasm rather than to the spindle apparatus.

To further evaluate developmental potential, we plated six ST blastocysts derived from vitrified spindles onto feeder cells and established two ESC lines (33%). We transplanted four ST blastocysts from vitrified spindles into a recipient that resulted in the timely birth of a healthy infant (Supplementary Fig. 8).

Postnatal development of monkey ST offspring

Although the technical feasibility of mtDNA replacement is documented for human embryos and ESCs, questions remain regarding whether a 'mismatch' between mtDNA and nuclear DNA haplotypes may cause mitochondrial dysfunctions in ST children[15]. To address these concerns, we conducted a 3-year follow-up study on monkey ST offspring born in 2009 (ref. 6).

The growth and development of four healthy infants following ST procedure was evaluated during the postnatal period (Supplementary Fig. 9a). Their overall health, including routine blood and body-weight measurements monitored from birth to 3 years were comparable to age-matched controls. The values for haemoglobin, red blood cell and white blood cell counts, mean corpuscular volume and haemoglobin concentrations were all within normal ranges (Supplementary Table 7). We also measured blood chemistry and arterial blood gas parameters and demonstrated that metabolic status of ST offspring is comparable to controls (Supplementary Table 7). In addition, the body-weight gain for the ST juvenile monkeys was similar to that of age-matched controls (Supplementary Fig. 9b). We also confirmed that ATP levels and mitochondrial membrane potential ($\Delta\Psi$m) in skin fibroblasts were similar to those of controls (Supplementary Fig. 10). Finally, there were no significant changes in mtDNA carryover and heteroplasmy in blood and skin samples with age (Supplementary Fig. 9c).

Discussion

This report summarizes our effort to test an mtDNA replacement in unfertilized human oocytes, initially developed and optimized in a monkey model. The results demonstrate that the ST procedure can be performed with high efficiency in human oocytes. Manipulated oocytes also supported high fertilization rates similar to those of controls. However, approximately half of the human ST zygotes had abnormal fertilization, primarily as a result of excessive pronuclear numbers. This was an unexpected outcome that was not observed with monkey oocytes. Our follow-up studies indicated that this is caused by the failure to complete meiosis and segregate chromosomes into the second PB, probably owing to premature activation. A set of haploid genetic material is normally discarded during asymmetrical cell division into the second PB, while the other half forms the female pronucleus. By genetic analysis of ESCs derived from abnormally fertilized zygote, we confirmed the triploid nature and presence of two sets of female chromosomes.

The spindle-chromosomal apparatus in MII oocytes is an extremely sensitive structure that can easily be perturbed by physical or chemical manipulations. Our initial attempts to isolate and transplant monkey MII spindles were unsuccessful owing to similar problems with spontaneous resumption of meiosis[6]. Procedures were optimized to avoid this negative outcome and current ST protocols allow maintenance of an intact MII spindle and normal fertilization. It seems that human MII oocytes are more sensitive to spindle manipulations and further improvements and optimizations will be required for future clinical applications. Maintenance of meiotic spindles in MII oocytes is dependent on the activity of M-phase-specific kinases including maturation-promoting factor (MPF) and mitogen-activated protein kinase[16]. Under normal conditions, sperm entry triggers degradation of kinase activities and chromosome segregation mediated by oscillations of intracellular Ca^{2+} concentrations[17]. However, an influx of calcium induced by mechanical or chemical manipulations can induce parthenogenetic activation of oocytes and resumption of meiosis[18]. Thus, ST manipulations in a medium without Ca^{2+} or supplementations with MG132 could potentially avoid problems with spontaneous activation[19,20].

Morphological evaluation of fertilization and early detection of abnormal pronuclear and/or polar body formation seems to be critical to separate normal and abnormal ST embryos. Blastocyst development and ESC isolation in normally fertilized ST zygotes were similar to controls. We also confirmed that all ESC lines derived from these ST embryos are karyotypically normal.

Two of the nine ESC lines (22%) derived from non-manipulated oocytes also showed chromosomal abnormalities. Because aberrations were confined to the sex chromosomes (47 XYY and 45 XO), it is possible that this was induced by sperm carrying either two Y chromosomes or no Y chromosome.

Despite the risk of abnormal pronuclear formation and aneuploidy in a portion of ST zygotes, embryo development and ESC isolation rates in normal ST zygotes are comparable to intact controls. Based on our estimates of retrieving on average 12 MII oocytes, 35% normal (2PN/2PB) fertilization rates, and 60% blastocyst development, at least two ST blastocysts suitable for transfers can be generated during a single cycle for each patient.

The safety of the ST procedure is also dependent on the amount of mutated mtDNA co-transferred with spindles. Importantly, mtDNA carryover in ST embryos and ESC lines is technically undetectable or below 1%. In most patients with mtDNA diseases, a threshold of 60% or higher of mutated mtDNA must be reached for clinical features to appear. Thus, it is unlikely that low mtDNA carryover during ST would cause disease in children. Segregation of mutated mtDNA to specific tissues during development and ageing may hypothetically result in a significant accumulation of the mutant load. However, analysis of mtDNA carryover in monkey ST offspring discovered

no detectable mtDNA segregation into different tissues[9]. In addition, there were no changes in heteroplasmy levels during postnatal development of monkeys. Thus, carryover, segregation and tissue-specific accumulation of mutant mtDNA molecules in ST children seem unlikely to be major concerns.

Birth of a healthy monkey infant after oocyte freezing marks an important milestone in applying the ST technology to patients. Transplantation of vitrified spindles into fresh cytoplasms yields the best results, comparable to controls. However, fertilization of vitrified cytoplasts even with fresh spindles was compromised. These remarkable findings indicate that the damage after cryopreservation is confined mainly to the eggs' cytoplasm, not to the chromosomes and spindles as commonly believed[21]. Our observations also reveal another unexpected potential clinical application of the ST technique, suggesting that spindles in sub-optimally cryopreserved oocytes can be rescued by transplanting into fresh cytoplasts.

Follow-up postnatal studies in four monkeys produced by ST provide convincing evidence that oocyte manipulation and mtDNA replacement procedures are compatible with normal development. These monkeys were derived by combining nuclear and mtDNA from the two genetically distant subpopulations of rhesus macaques. Mitochondrial and nuclear genetic differences between these monkeys are considered to be as distant as those between some different primate species[22], thus imitating haplotype differences between humans. Concerns have been raised that nuclear and mtDNA incompatibilities between mtDNA patients and cytoplast donors may cause a 'mismatch' and mitochondrial dysfunctions in ST children even in the absence of mutations[15]. On the basis of our long-term observations, it is reasonable to speculate that nuclear–mtDNA interactions are conserved within species.

Pioneering work in nonhuman primates is critical for the development, and safety and efficacy evaluations, of new treatments[23,24]. It is important that scientists and clinicians further optimize ST protocols for human oocytes and ensure that these procedures are safe. It is also crucial that the US Food and Drug Administration initiates careful review of these new developments. Such oversight will be important to establish safety and efficacy requirements and guide clinical trials. Current US National Institutes of Health funding restrictions surrounding these innovative reproductive technologies will also require amendments to support federally funded clinical trials.

METHODS SUMMARY

The study protocols were approved by both the Oregon Health & Science University Embryonic Stem Cell Research Oversight Committee and the Institutional Review Board.

Mature oocytes were donated by volunteers and ST procedures were carried out as described[5,6]. Oocytes were fertilized, cultured to blastocysts and used for ESC isolation. Detailed methods are described in Supplementary Information at www.nature.com/nature.

Received 9 June; accepted 3 October 2012.
Published online 24 October 2012.

1. Gropman, A. L. Diagnosis and treatment of childhood mitochondrial diseases. *Curr. Neurol. Neurosci. Rep.* **1,** 185–194 (2001).
2. Haas, R. H. *et al.* Mitochondrial disease: a practical approach for primary care physicians. *Pediatrics* **120,** 1326–1333 (2007).
3. Schaefer, A. M. *et al.* Prevalence of mitochondrial DNA disease in adults. *Ann. Neurol.* **63,** 35–39 (2008).
4. Elliott, H. R., Samuels, D. C., Eden, J. A., Relton, C. L. & Chinnery, P. F. Pathogenic mitochondrial DNA mutations are common in the general population. *Am. J. Hum. Genet.* **83,** 254–260 (2008).
5. Tachibana, M., Sparman, M. & Mitalipov, S. Chromosome transfer in mature oocytes. *Nature Protocols* **5,** 1138–1147 (2010).
6. Tachibana, M. *et al.* Mitochondrial gene replacement in primate offspring and embryonic stem cells. *Nature* **461,** 367–372 (2009).
7. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N. Engl. J. Med.* **350,** 1353–1356 (2004).
8. Danan, C. *et al.* Evaluation of parental mitochondrial inheritance in neonates born after intracytoplasmic sperm injection. *Am. J. Hum. Genet.* **65,** 463–473 (1999).
9. Lee, H.-S. *et al.* Rapid mitochondrial DNA segregation in primate preimplantation embryos precedes somatic and germline bottleneck. *Cell Reports* **1,** 506–515 (2012).
10. Burgstaller, J. P., Schinogl, P., Dinnyes, A., Muller, M. & Steinborn, R. Mitochondrial DNA heteroplasmy in ovine fetuses and sheep cloned by somatic cell nuclear transfer. *BMC Dev. Biol.* **7,** 141 (2007).
11. Wakayama, T. & Yanagimachi, R. The first polar body can be used for the production of normal offspring in mice. *Biol. Reprod.* **59,** 100–104 (1998).
12. Susko-Parrish, J. L., Leibfried-Rutledge, M. L., Northey, D. L., Schutzkus, V. & First, N. L. Inhibition of protein kinases after an induced calcium transient causes transition of bovine oocytes to embryonic cycles without meiotic completion. *Dev. Biol.* **166,** 729–739 (1994).
13. Forman, E. J. *et al.* Oocyte vitrification does not increase the risk of embryonic aneuploidy or diminish the implantation potential of blastocysts created after intracytoplasmic sperm injection: a novel, paired randomized controlled trial using DNA fingerprinting. *Fertil. Steril.* **98,** 644–649 (2012).
14. Rienzi, L. *et al.* Consistent and predictable delivery rates after oocyte vitrification: an observational longitudinal cohort multicentric study. *Hum. Reprod.* **27,** 1606–1612 (2012).
15. Moreno-Loshuertos, R. *et al.* Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants. *Nature Genet.* **38,** 1261–1268 (2006).
16. Fisher, D. L., Brassac, T., Galas, S. & Doree, M. Dissociation of MAP kinase activation and MPF activation in hormone-stimulated maturation of *Xenopus* oocytes. *Development* **126,** 4537–4546 (1999).
17. Runft, L. L., Jaffe, L. A. & Mehlmann, L. M. Egg activation at fertilization: where it all begins. *Dev. Biol.* **245,** 237–254 (2002).
18. Mitalipov, S. M., Nusser, K. D. & Wolf, D. P. Parthenogenetic activation of rhesus monkey oocytes and reconstructed embryos. *Biol. Reprod.* **65,** 253–259 (2001).
19. Gao, S., Han, Z., Kihara, M., Adashi, E. & Latham, K. E. Protease inhibitor MG132 in cloning: no end to the nightmare. *Trends Biotechnol.* **23,** 66–68 (2005).
20. Kikuchi, K. *et al.* Maturation/M-phase promoting factor: a regulator of aging in porcine oocytes. *Biol. Reprod.* **63,** 715–722 (2000).
21. Mandelbaum, J. *et al.* Effects of cryopreservation on the meiotic spindle of human oocytes. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **113** (Suppl 1), S17–S23 (2004).
22. Smith, D. G. Genetic characterization of Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*). *Comp. Med.* **55,** 227–230 (2005).
23. Donnez, J. *et al.* Children born after autotransplantation of cryopreserved ovarian tissue. a review of 13 live births. *Ann. Med.* **43,** 437–450 (2011).
24. Lee, D. M. *et al.* Live birth after ovarian tissue transplant. *Nature* **428,** 137–138 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to acknowledge the Oregon Health & Science University (OHSU) Embryonic Stem Cell Research Oversight Committee and the Institutional Review Board for providing oversight and guidance. We thank oocyte and sperm donors and staff at the Women's Health Research Unit at the Center for Women's Health, University Fertility Consultants and Reproductive Endocrinology & Infertility Division at the Department of Obstetrics & Gynecology of Oregon Health & Science University for their support and procurement of human gametes. The Division of Animal Resources, Surgery Team, Assisted Reproductive Technology & Embryonic Stem Cell Core, Endocrine Technology Core, Imaging & Morphology Core, Flow Cytometry Core and Molecular & Cellular Biology Core at the Oregon National Primate Research Center provided expertise and services for the nonhuman primate research. Hamilton Thorne Inc., donated an XYClone laser system for this study. We are grateful to W. Sanger and D. Zaleski for karyotyping services, C. Penedo for microsatellite analysis and J. Hennebold for consulting on metabolic assays. We are also indebted to A. Steele, R. Cervera Juanes and E. Wolff for their technical support. The human oocyte/embryo research was supported by grants from the OHSU Center for Women's Health Circle of Giving and other OHSU institutional funds, as well as the Leducq Foundation. The nonhuman primate study was supported by grants from the National Institutes of Health HD063276, HD057121, HD059946, EY021214 and 8P51OD011092.

Author Contributions M.T., P.A., J.J. and S.M. conceived the study, designed experiments and wrote institutional review board protocols. P.A., M.S. and N.M.G. coordinated recruitment of participants. P.A., K.M., D.B., D.L., D.W. and P.P. performed ovarian stimulation and oocyte recovery. M.T. conducted ST micromanipulations. M.S., K.M. and S.M. performed ICSI. M.T., M.S., J.W., D.M.S., N.M.G., R.T.-H. and E.K. conducted ESC derivation and characterization. S.G. analysed teratoma tumours. M.T., H.M. and D.M.S. performed DNA/RNA isolations, metabolic and mtDNA analyses. C.R., M.T., M.S.,H.-S.L., R.S. and S.M. conducted monkey studies. M.T., R.S., J.J., P.P. and S.M. analysed data and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M. (mitalipo@ohsu.edu).

ARTICLE

doi:10.1038/nature11800

Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants

Daniel Paull[1], Valentina Emmanuele[2], Keren A. Weiss[1], Nathan Treff[3], Latoya Stewart[1], Haiqing Hua[1,4], Matthew Zimmer[1], David J. Kahler[1], Robin S. Goland[4], Scott A. Noggle[1], Robert Prosser[5], Michio Hirano[2], Mark V. Sauer[5,6]* & Dieter Egli[1]*

Mitochondrial DNA mutations transmitted maternally within the oocyte cytoplasm often cause life-threatening disorders. Here we explore the use of nuclear genome transfer between unfertilized oocytes of two donors to prevent the transmission of mitochondrial mutations. Nuclear genome transfer did not reduce developmental efficiency to the blastocyst stage, and genome integrity was maintained provided that spontaneous oocyte activation was avoided through the transfer of incompletely assembled spindle–chromosome complexes. Mitochondrial DNA transferred with the nuclear genome was initially detected at levels below 1%, decreasing in blastocysts and stem-cell lines to undetectable levels, and remained undetectable after passaging for more than one year, clonal expansion, differentiation into neurons, cardiomyocytes or β-cells, and after cellular reprogramming. Stem cells and differentiated cells had mitochondrial respiratory chain enzyme activities and oxygen consumption rates indistinguishable from controls. These results demonstrate the potential of nuclear genome transfer to prevent the transmission of mitochondrial disorders in humans.

A crucial determinant of the phenotypic severity in most maternally inherited mitochondrial diseases is heteroplasmy, that is, the proportion of mutant, relative to total, mitochondrial DNA (mtDNA) in a cell. Owing to the cytoplasmic segregation of mitochondria during cell division, the level of heteroplasmy is subject to broad fluctuations, in particular during the developmental expansion of mtDNA from the premeiotic germ cell to the mature human oocyte[1–4]. As a result, an unaffected carrier of a mtDNA mutation may have an affected child. Although prenatal genetic diagnosis can select embryos with a reduced mutation load, variation between blastomeres in single embryos limits the effectiveness of such screening[3], and considerable levels of mutant mtDNA can remain, resulting in a carrier[5].

On the basis of these considerations, the Nuffield Council on Bioethics has endorsed research to prevent transmission of mtDNA mutations[6], including the transfer of the nuclear genome into an enucleated oocyte containing normal mitochondria. In mice, transfer between fertilized eggs (zygotes) is effective in preventing the transmission of pathogenic mtDNA[7] and in rhesus monkeys, genome exchange between unfertilized oocytes gave rise to live births[8]. In human cells, the transfer of pronuclei between zygotes resulted in minimal carryover of donor mtDNA[9]. However, the exchange reduced developmental potential, possibly because the transfer introduced an abnormal centrosome number, resulting in multipolar spindles[10] and aneuploidy[11]. As the centrosome is sperm-derived[12], nuclear genome exchange before fertilization avoids this issue. Furthermore, pronuclear transfer requires the fertilization of both donor and recipient oocytes, resulting in the destruction of half of the embryos. By contrast, human oocytes would only be fertilized after successful genome exchange.

To determine the consequences of nuclear genome transfer in unfertilized oocytes, we chose parthenogenetic activation instead of fertilization, as it avoids the generation of human embryos for research (Supplementary Fig. 1). Transfer of the nuclear DNA between oocytes of two unrelated donors resulted in the exchange of the mitochondrial genotype and the elimination of mtDNA variants, including a variant found in the *MT-TV* gene (encoding mt-tRNA^{Val}). However, nuclear genome transfer frequently induced premature oocyte activation and failure to extrude the second polar body normally. As manipulation-induced karyotypic abnormalities present a risk beyond those normally incurred during assisted reproductive technologies, this is probably an obstacle to the clinical translation of this technique. We found that premature activation could be prevented by partial depolymerization of the spindle–chromosome complex through cryopreservation or cooling to room temperature, allowing normal polar body extrusion, efficient development to the blastocyst stage, and the derivation of karyotypically normal stem cells. Therefore, nuclear genome transfer, rather than the transfer of intact spindle–chromosome complexes, should be effective in preventing the transmission of mtDNA mutations.

Efficient development after genome exchange

We first determined that artificial activation of unfertilized oocytes resulted in extrusion of the second polar body (Supplementary Fig. 2a) and efficient development to the blastocyst stage (Fig. 1a), allowing derivation of four stem-cell lines (parthenogenetic embryonic stem (pES) cells 2–5 (pES2–5); Supplementary Fig. 2b–o). If polar body extrusion had accurately segregated sister chromatids, a haploid zygote with 23 chromosomes should result in stem-cell lines devoid of heterozygosity. Short tandem repeat (STR) genotyping and Affymetrix single nucleotide polymorphism (SNP) arrays showed that four out of four stem-cell lines were homozygous for all chromosomes (Supplementary Fig. 2p and Supplementary Table 1). All cell lines were diploid (Supplementary Fig. 2q–t), suggesting that the genome had undergone endoreplication, as previously observed in mouse parthenotes[13]. These results suggest that parthenogenesis is appropriate for *in vitro* studies on the feasibility and consequences of oocyte genome exchange.

To exchange the nuclear genome between oocytes of two different donors, menstrual cycles were synchronized using oral contraceptives, with synchronized retrieval successful in all (four out of four)

[1]The New York Stem Cell Foundation Laboratory, New York 10032, USA. [2]Department of Neurology, Columbia University, New York 10032, USA. [3]Reproductive Medicine Associates of New Jersey, New Jersey 07960, USA. [4]Naomi Berrie Diabetes Center, College of Physicians and Surgeons, Columbia University, New York 10032, USA. [5]Center for Women's Reproductive Care, College of Physicians and Surgeons, Columbia University, New York 10019, USA. [6]Department of Obstetrics and Gynecology, College of Physicians and Surgeons, Columbia University, New York 10032, USA.
*These authors contributed equally to this work.

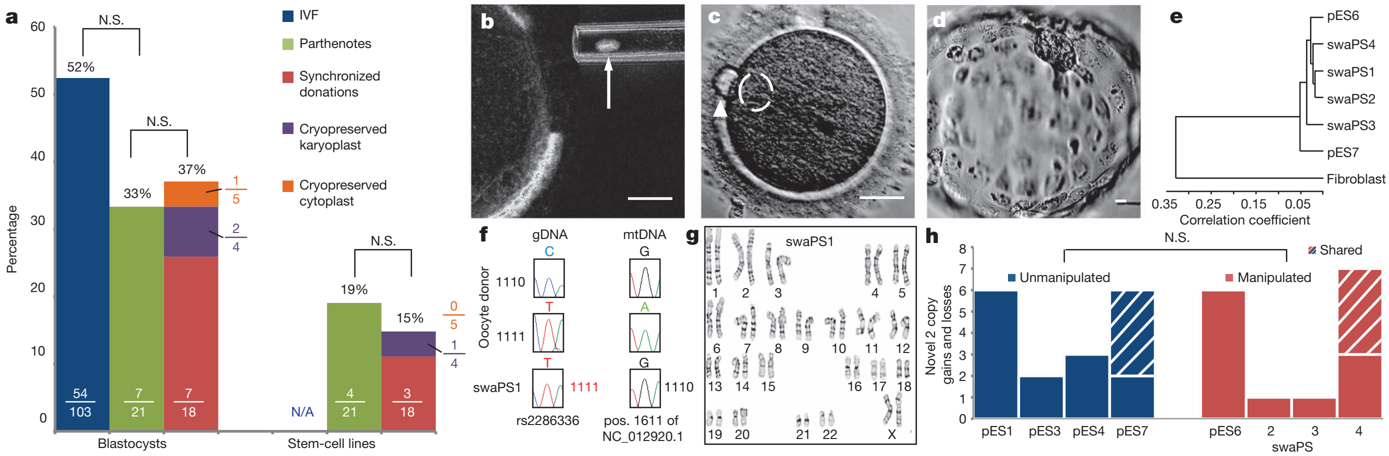


Figure 1 | Efficient development and genomic integrity after nuclear genome exchange. **a**, Developmental potential of IVF embryos, parthenogenic and genome-exchanged oocytes. Numbers in each bar displays the proportions of oocytes developing to the indicated stage. **b**, Removal of the karyoplast (indicated by arrow). **c**. Polar body extrusion (arrowhead) and pronuclear formation (dashed circle) at 4 h after activation **d**, Blastocyst stage at day 7 after activation. **e**, Cluster diagram of global gene expression profiles of swaPS cell lines. **f**, Sanger sequences of genomic DNA (gDNA) and mtDNA. **g**, Karyotype of swaPS1 cells at P3. **h**, Copy number variation analysis. N.S., not significant. Scale bars, 25 μm for all panels.

women. Four oocyte donors donated a total of 62 mature metaphase II (MII) oocytes (19, 17, 11 and 15 oocytes for the individual donors). Using microtubule birefringence, 18 of these oocytes had the nuclear genome removed (Fig. 1b) and subsequently fused to enucleated oocytes using either Sendai virus or an electrical pulse. Development after exchange was very efficient: of the 18 oocytes, seven developed to the blastocyst stage, with at least one blastocyst for each donor (Fig. 1a, d and Supplementary Fig. 3a–h). From these blastocysts, three swapped pluripotent stem (swaPS) cell lines were derived. swaPS cells expressed markers of pluripotency, had gene expression profiles comparable to those of established embryonic and parthenogenic embryonic stem-cell lines, and were able to differentiate into cell types and tissues of each germ layer (Fig. 1e and Supplementary Fig. 3i–t). Sequencing of mtDNA and nuclear DNA polymorphisms and STR genotyping confirmed that nuclear genome transfer had resulted in the exchange of the mitochondrial genotype (Fig. 1f, Supplementary Fig. 4a and Supplementary Tables 2 and 3).

To determine whether the transfer affected the integrity of the nuclear genome, karyotype analysis and high-resolution SNP arrays were used. Both swaPS1 and swaPS2 cells had normal karyotypes of 46,XX chromosomes, whereas swaPS3 cells contained an extra chromosome 12 (Fig. 1g and Supplementary Fig. 3u, v). The origin of this additional chromosome was due to a mitotic segregation error, as chromosome 12 did not contain regions of heterozygosity by SNP array analysis (Supplementary Information). Although trisomy 12 is a frequent artefact of *in vitro* stem-cell culture[14], karyotyping of swaPS1 cells after 9 months in culture revealed no alterations (Supplementary Information). Manipulation-induced strain on chromosomes during genome exchange might result in chromosome breaks and faulty repair, as no homologous template is available after segregation of sister chromatids. Duplication during S phase would result in homozygous copy number variants (CNVs). We identified an average of 3.75 homozygous CNVs in five cell lines with spindle transfer and 4.25 in four unmanipulated parthenogenetic stem-cell lines (Fig. 1h). swaPS4 cells and the unmanipulated cell line pES7 originate from the same donor, and were found to share more than half of the CNVs, suggesting that most, if not all, CNVs originated in the female germ line, and were not generated by the manipulation. We also found no differences in the number of heterozygous CNVs (Supplementary Fig. 3x). To determine whether epigenetic alterations resulted in gene expression changes after genome transfer, gene expression in manipulated cell lines was compared to unmanipulated controls. Only one gene (*THBS1*, thrombospondin 1) was expressed at significantly lower levels ($P < 0.01$), with no genes having increased expression.

swaPS2 and swaPS3 cells were homozygous for 99.8% of the genome, consistent with accurate extrusion of a haploid genome into the polar body. The 0.2% of heterozygosity was due to CNVs that diverged in sequence (Supplementary Information). By contrast, the swaPS1 cell line was homozygous for merely 43.3% of the genome (Fig. 2a and Supplementary Table 4), consistent with a cell line that had undergone the first, but not the second, meiosis[15]. To determine the frequency of chromosome segregation errors during the preimplantation stages, we biopsied polar bodies and blastomeres, or used morulas and blastocysts, for analysis. Using whole-genome amplification, followed by either PCR of loci on all 23 chromosomes or microarray-based analysis, we found that seven out of nine preimplantation embryos were homozygous throughout the genome, demonstrating normal polar body extrusion (Supplementary Fig. 4b–e). One was heterozygous for all chromosomes, whereas another showed a copy number gain and heterozygosity on chromosome 4, demonstrating that polar body extrusion had been inaccurate (Fig. 2b). Failed, or inaccurate extrusion of the polar body was probably due to the transfer procedure, as we did not observe any heterozygosity in unmanipulated stem-cell lines[16] (Supplementary Fig. 2p).

Immature spindles prevent spontaneous activation

Through transfer of karyoplasts into oocytes of either the same or a different donor, we determined that after transfer using electrical pulses, all oocytes (13 out of 13) formed one or two pronuclei 3–5 h after transfer, suggesting that the electrical pulse had caused premature exit from meiosis. Parthenotes with molecularly confirmed karyotypic abnormalities were all derived from oocytes that were activated as a result of the transfer. By contrast, none of the oocytes (0 out of 14) fused to karyoplasts using Sendai virus activated as a result of the manipulation. All remained at meiosis for 4–6 h after transfer, and extruded a second polar body only after artificial activation using a calcium ionophore followed by incubation in the translation inhibitor puromycin (Fig. 2c, d).

We reasoned that manipulation-induced activation was related to the spindle–chromosome complex. When oocytes without a spindle were exposed to an identical electrical fusion pulse during somatic cell nuclear transfer, they remained stable in meiosis[16]. Furthermore, when sperm is injected prematurely, within an hour after extrusion of the first polar body and before the assembly of a mature MII spindle, most oocytes fail to activate[17]. This suggests that after an activating stimulus, only mature spindles with bipolar attachment of chromosomes generate a signal promoting the exit from meiosis, thereby ensuring accurate chromosome segregation. To explore

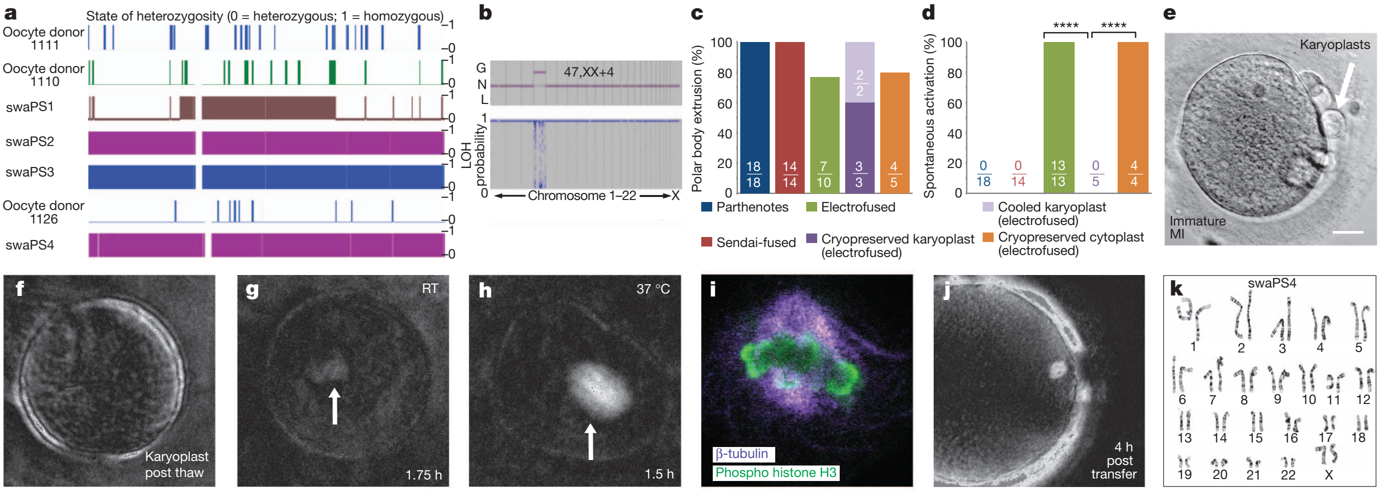


Figure 2 | Spontaneous activation can be prevented through spindle cooling. **a**, Affymetrix SNP microarray (chromosome 7). **b**, Array analysis at the cleavage stage, revealing trisomy on chromosome 4. G, gain; L, loss; LOH, loss of heterozygosity; N, normal. **c**, Frequency of polar body extrusion. **d**, Frequency of manipulation-induced activation. $^{****}P < 0.0001$. **e**, Karyoplasts after cryopreservation. Scale bar, 25 μm. **f–h**, Spindle birefringence at indicated temperatures and time points post thaw. Arrows indicate site of spindle. RT, room temperature. **i**, Confocal analysis of a thawed karyoplast after 2 h at room temperature. **j**, Spindle birefringence after transfer of a thawed karyoplast into an enucleated oocyte. **k**, Karyotype of stem-cell line derived from a cryopreserved karyoplast, swaPS4, at P4.

whether immature spindles prevented manipulation-induced activation, we monitored an oocyte progressing from metaphase I (MI) to MII and aspirated the nuclear genome within 1 h after extrusion of the first polar body, while the spindle was still immature (Supplementary Fig. 5a). Transfer-induced activation did not occur and only after artificial activation did the oocyte extrude the polar body and develop to the blastocyst stage, allowing derivation of a pluripotent stem-cell line (pES6) with a normal karyotype and homozygosity across the genome (Supplementary Fig. 5b–g and Supplementary Table 5).

As timing the removal of the nuclear genome relative to the extrusion of the first polar body may not be practical, we used reduced temperatures to induce partial spindle depolymerization in mature MII oocytes. As the removal of the spindle–chromosome complex relies on the presence of a birefringent spindle, we extracted the nuclear genome on the heated stage of a microscope, before exposing the karyoplast to reduced temperatures. Two karyoplasts were placed on ice for 2 h, and then re-transferred by electrofusion into enucleated oocytes kept at 37 °C. Fused oocytes remained in meiosis for more than 3 h post transfer and required an artificial activation stimulus for polar body extrusion. We next extracted karyoplasts from 30 oocytes and cryopreserved them below the zona pellucida of an immature oocyte (Fig. 2e). After thawing, 27 out of 30 karyoplasts were intact. Spindle birefringence was not detected immediately after thaw but re-formed in karyoplasts kept at 37 °C for approximately 2 h (two out of two), but not when maintained at room temperature (none out of three) (Fig. 2f–h). Cryopreservation did not result in the dispersion of chromosomes: in all five karyoplasts, chromosomes remained attached to microtubules (Fig. 2i), consistent with the finding that some, but not all spindle microtubules are cold sensitive[18]. Within 1 h of thawing, three karyoplasts were fused to enucleated oocytes of a different donor using electrical pulses. At 4 h after transfer, we observed a birefringent spindle with perpendicular alignment to the oolemma (Fig. 2j). All (three out of three) oocytes remained stable at meiosis and extruded the second polar body only after artificial activation, forming a single pronucleus (Supplementary Fig. 6a, b). Two of the three parthenotes (66%) developed to the blastocyst stage, allowing the derivation of the karyotypically normal stem-cell line swaPS4 (Fig. 2k and Supplementary Fig. 6c–l). Homozygosity throughout the genome showed that polar body extrusion had normally segregated sister chromatids (Fig. 2a and Supplementary Tables 4 and 5). An additional thawed karyoplast was kept at 37 °C for 4 h, and only then fused to the oocyte, resulting in spontaneous activation and a failure to extrude the second polar body.

Furthermore, we vitrified MII oocytes. Immediately after thaw, birefringence of the microtubule spindle was undetectable in all oocytes (none out of six). After incubation for 1–2 h at 37 °C, birefringence became visible in all oocytes, which were subsequently enucleated. Spindle–chromosome complexes from oocytes of a different donor retrieved on the day of thawing were transferred into five enucleated oocytes by electrofusion, with one karyoplast failing to fuse. One of the five oocytes developed to the blastocyst stage (Fig. 1a and Supplementary Fig. 6m–o). Unlike in oocytes after transfer of cryopreserved karyoplasts, oocytes underwent spontaneous activation (four out of four monitored during the relevant time period, Fig. 2d). These results demonstrate that the maturation of the spindle–chromosome complex has a major role in the ability of human oocytes to exit meiosis.

Stable exchange of mitochondrial genotypes

To determine mitochondrial genotypes, we identified polymorphic mtDNA variants (SNPs) in the hypervariable regions (HVR1 and HVR2) of each donor (Supplementary Table 6), and sequenced the complete mtDNA genome of two donors (Supplementary Table 7). Two polymorphisms, m.4715A>G (*MT-ND2*) and m.16129A>G (non-coding region), were homoplasmic by restriction fragment length polymorphism (RFLP). Moreover, one of the donors was homoplasmic for a variant at the 3′ terminus of the *MT-TV* gene (m.1670A>T) that has been identified as a rare polymorphism[19]. Although no differences in the mt-tRNA^{Val} steady-state level were observed in cultured fibroblasts, a functional significance is possible as an analogous 3′ terminus polymorphism in mt-tRNA^{Glu} has been identified as a pathogenic mutation[20] (Supplementary Fig. 7a–d).

To estimate the amount of potential mtDNA carryover, we quantified mtDNA in human oocytes and karyoplasts using quantitative PCR (qPCR). Karyoplasts had an average mtDNA copy number of 1,129 ± 785 (mean ± s.d., $n = 22$), or 0.36% of the total mtDNA found within MII oocytes (311,146 ± 206,521, $n = 5$) (Fig. 3a). This corresponded with volumetric measurements of karyoplasts, which were 0.89% of that of intact oocytes (4,961 ± 1,964 μm^3 versus 559,093 ± 245,217 μm^3, respectively; $n = 18$ and 21) (Fig. 3a). Staining of mitochondria in oocytes and karyoplasts with MitoTracker or antibodies recognizing the complex V α-subunit indicated that the

spindle–chromosome complex was devoid of mitochondria (Fig. 3b, c). We therefore expected a mtDNA carryover of less than 1%.

Heteroplasmy was determined by last-hot cycle PCR RFLP, with a detection threshold of approximately 2%, and allele refractory mutation system (ARMS)-qPCR, with a detection threshold of approximately 0.5% (refs 21, 22; Supplementary Figs 7 and 8). As considerable fluctuations in heteroplasmy have been reported in blastomeres of mouse and monkey embryos[23,24], we first quantified 17 embryos at the cleavage stage, including individual blastomeres, and found heteroplasmy to be less than 0.5%. In seven samples that reached morula or blastocyst stage, the average heteroplasmy remained less than 0.5%, and in all cases 1% or lower. Overall, preimplantation embryos had a mean heteroplasmy of 0.31% ± 0.27% ($n = 24$; Fig. 3d and Supplementary Fig. 8e).

With the generation of stem-cell lines, we asked whether the original mitochondrial genotype could re-emerge after extensive passaging, clonal expansion, cellular differentiation and reprogramming. Quantification at passages (P)2–P59 (1 year in culture), showed that heteroplasmy in the swaPS1, swaPS2 and swaPS4 cell lines remained undetectable (Fig. 3d, e and Supplementary Figs 7e–l and 8f). mtDNA heteroplasmy was detected at low levels in swaPS3 cells at P4–P10 (2.79% ± 0.27%; range: 2.3–3.1% by RFLP), but became undetectable from P14 onward (Fig. 3f and Supplementary Fig. 7e, i–j). As heterogeneity in heteroplasmy may not be detected by population analysis, ten colonies for each swaPS cell line were grown from single cells. In all 40 colonies, heteroplasmy was undetectable (Fig. 3d and Supplementary Fig. 8f).

In mice, specific mitochondrial genotypes can selectively expand in differentiated cells, resulting in altered levels of heteroplasmy[25,26]. To determine whether such alterations occurred after differentiation, cells of each germ layer (pancreatic cells, neurons, fibroblasts and cardiomyocytes) were differentiated *in vitro* (Fig. 3g–i and Supplementary Video 1). Heteroplasmy was undetectable by either ARMS-qPCR or RFLP (Fig. 3d, e and Supplementary Fig. 8g).

A bottleneck in the mitotic inheritance of mitochondrial DNA mutations occurs during induced pluripotent stem (iPS) cell generation, resulting in iPS cell colonies with differing percentages of a mitochondrial mutation[27]. To test whether such a bottleneck (mimicking the one in the female germ line) could alter the ratio of the two mitochondrial genotypes, we reprogrammed fibroblasts differentiated from swaPS cells (Fig. 4a) into iPS cells (Supplementary Fig. 9a–h). In 43 iPS cell colonies and 6 cell lines, termed swiPS, heteroplasmy was undetectable (mean heteroplasmy 0.01% ± 0.04%; $n = 43$; Fig. 3d and Supplementary Fig. 9i). As none of these manipulations resulted in re-emergence of significant levels of heteroplasmy, we next asked whether swaPS cells retained any mitochondrial DNA transferred with the karyoplast. Using repeat-PCR amplification of a product in the HVR, with five SNPs per primer pair, we could specifically amplify the minority product. In neither swaPS1 nor swaPS2 cells could the minority product be detected beyond P2 (detection limit 0.0001%), whereas in swaPS3 cells detection was seen until P18, after which it was undetectable (detection limit 0.1%; Supplementary Fig. 9j, k). As stem cells and fibroblasts contain less than 1,500 mtDNA copies[28], a considerable proportion of the cell population must be homoplasmic.

Normal mitochondrial activity in swaPS cells

Despite the low levels of mitochondrial heteroplasmy, we considered the possibility that the presence of two mitochondrial genotypes within the same cell might impair mitochondrial functions[26]. We determined that swaPS1 and swaPS2 cells had population doubling times of 33 and 31.5 h, respectively, a proliferation rate comparable to embryonic stem (ES) cell lines[29], suggesting no metabolic disadvantage as a result of the exchange. We next determined the biochemical

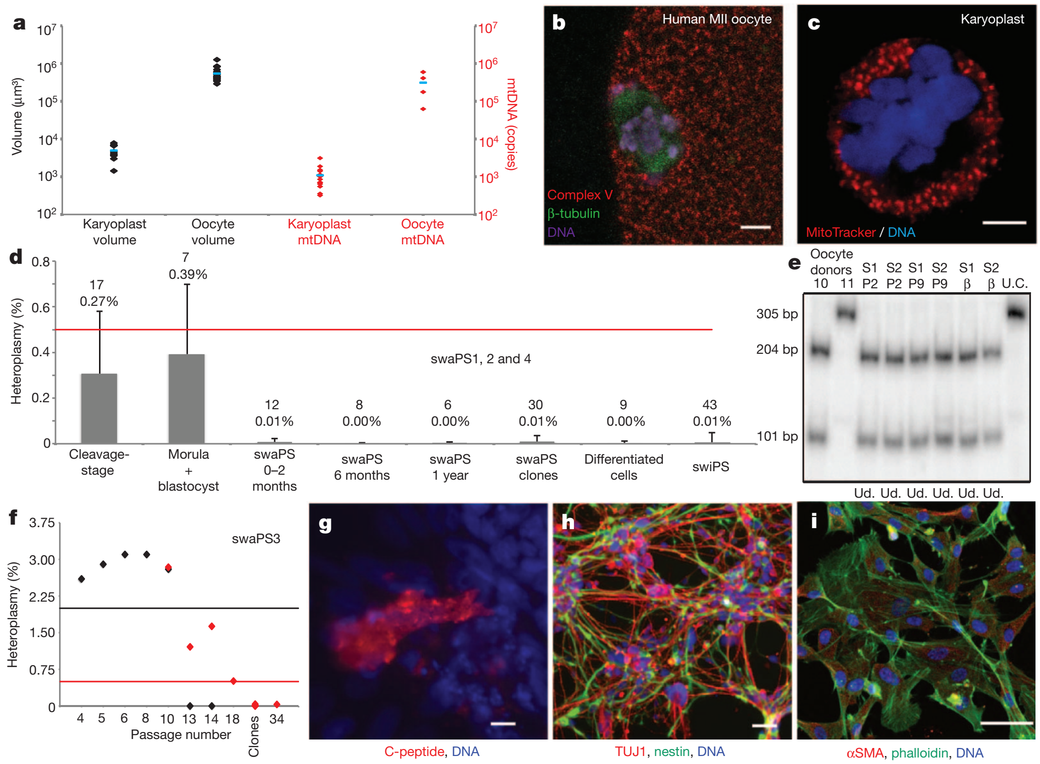


Figure 3 | Low levels of mtDNA carryover. **a**, Volume (black) and mtDNA copies (red) of karyoplasts and oocytes. Blue bars denote the mean. **b**, **c**, Distribution of mitochondria in the oocyte (**b**) and the karyoplast (**c**). **d**, Mean heteroplasmy quantification by ARMS-qPCR. Red line indicates limit of detection. Error bars indicate s.d., with the mean value and n number shown. **e**, RFLP analysis of swaPS1 and swaPS2 (S1 and S2, respectively) at P2 and P9 and as β-cells. bp, base pairs; U.C., undigested control; Ud., undetectable. **f**, Heteroplasmy in swaPS3 cells. ARMS-qPCR (red diamonds) and RFLP (black diamonds); black and red lines indicate detection limits. **g**–**i**, Directed differentiation into β-cells (**g**), neurons (**h**) and fibroblasts (**i**). αSMA, α-smooth muscle actin. Scale bars, 5 μm (**b**, **c**) and 50 μm (**g**–**i**).

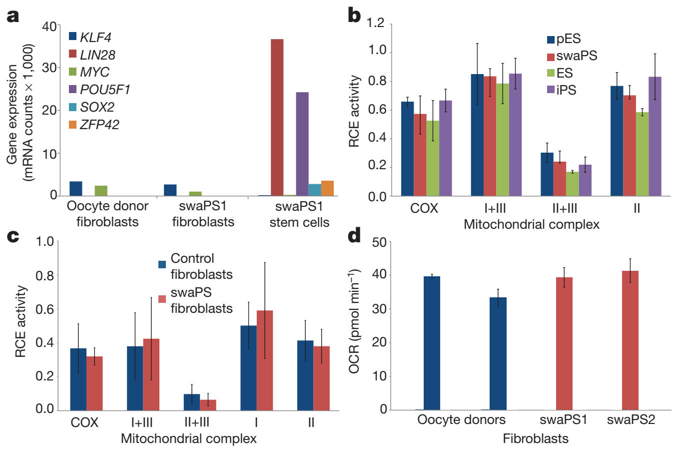


Figure 4 | swaPS cells support a normal metabolic profile. **a**, Nanostring gene expression analysis of fibroblasts derived from swaPS1 cells. mRNA counts per 100 ng RNA. **b**, Respiratory chain enzyme (RCE) activities in ES, pES, swaPS and iPS cell lines derived from oocyte donor skin cells. COX, cytochrome *c* oxidase (also known as complex IV). **c**, RCE activities of mitochondrial respiratory complexes in swaPS1- and swaPS2-derived fibroblasts (swaPS fibroblasts) compared with control fibroblasts. **d**, Analysis of basal oxygen consumption rate (OCR) in swaPS1- and swaPS2-derived fibroblasts compared with control fibroblasts. Error bars denote s.d.

activities of mitochondrial respiratory chain enzymes. No differences were found in comparisons between undifferentiated swaPS1 and swaPS2, ES, pES and iPS cell lines generated from oocyte donor skin fibroblasts (Fig. 4b). We also assessed mitochondrial respiratory chain enzyme activities in differentiated cell types (Fig. 4c) and basal oxygen consumption in live culture (Fig. 4d), and found no differences between swaPS-derived fibroblasts and oocyte donor skin fibroblasts.

Discussion

Here we show that the transfer of nuclear genomes into human oocytes results in efficient preimplantation development, with essentially complete and mitotically stable exchange of the mitochondrial lineage, supporting normal mitochondrial activity. The low levels of heteroplasmy that are detected immediately after transfer become undetectable and do not increase after clonal expansion, cellular differentiation or when exposed to a bottleneck. Any potentially remaining heteroplasmy is far below the levels required for clinical presentation of a mitochondrial mutation, and below the levels of heteroplasmy able to cause behavioural defects in mice[26]. Importantly, the transfer is compatible with the integrity of the nuclear genome, as it did not result in an increase in CNVs.

We found that manipulation of an intact spindle–chromosome complex frequently induced premature activation of the oocyte, resulting in karyotype abnormalities. Preventing such abnormalities is crucial for successful clinical application of nuclear genome transfer, as it is unclear whether visual inspection or even molecular assays can reliably eliminate abnormal embryos. Exposure of the karyoplast, but not of the cytoplast, to low temperatures prevented this manipulation-induced activation. Spindle–chromosome complexes at or below room temperature showed decreased birefringence of microtubules, a measure for the regularity of microtubule alignment. Importantly, after fusion and incubation at 37 °C, spindle birefringence returned, allowing polar body extrusion, normal preimplantation development, and derivation of karyotypically normal stem cells. Such reversible destabilization of spindle microtubules by low temperatures has been previously observed in human and animal oocytes[30,31]. During vitrification of human oocytes, partial depolymerization of the spindle occurs owing to the exposure to room temperature, rather than the vitrification itself[32], but does not result in the dispersion of chromosomes and does not increase the incidence of karyotypic abnormalities[33,34], allowing the adoption for clinical use[35]. In a concurrent study using temperatures of 37 °C to maintain a fully assembled spindle during manipulation[36,37], frequent signs of premature activation were found, despite the use of Sendai virus for fusion, which is less prone to induce activation than an electrical pulse[8]. As a result, more than half of the oocytes were lost due to a failure in polar body extrusion and the formation of karyotypically abnormal embryos. By contrast, during our study, all oocytes were briefly exposed to room temperature (21 °C), and when karyoplasts were transferred using Sendai virus, all oocytes remained at meiosis and extruded the polar body only after artificial activation. A further difference between the two studies is the timing of oocyte retrieval at 35 instead of 36 h after hormonal induction of ovulation. Both exposure to room temperature and early retrieval probably result in more oocytes with immature spindles at the time of manipulation. Bipolar attachment of chromosomes to spindle microtubules reduces the activity of auroraB kinase and of the spindle assembly checkpoint in a tension-dependent manner[38]. Manipulation may result in increased tension and decreased kinase target phosphorylation in fully assembled spindles, but not in spindles with incomplete bipolar attachment of chromosomes, thereby avoiding premature activation.

Although concerns have been raised that the transfer of the maternal genome between oocytes may introduce ‘epigenetic’ changes[39], we did not observe meaningful differences in gene expression between genome-exchanged and unmanipulated cells. Although our methods allow maternal nuclear genome transfer without apparent adverse consequences to the embryo, the accuracy of chromosome segregation at the second meiosis will need to be determined in a large number of samples. Furthermore, animal models may be required to determine whether the transfer of incompletely assembled spindles into oocytes is compatible with normal fertilization and development to term. An analogous technique for the transfer of the paternal genome by intracytoplasmic sperm injection is performed in approximately half of all *in vitro* fertilization (IVF) cycles in the United States[40], providing a precedent for clinical translation. Before proceeding with human clinical trials on the transfer of the maternal genome, it will be important to publicly discuss patient needs, ethical considerations, and to establish appropriate guidelines for the use of oocyte nuclear genome transfer in assisted reproduction.

METHODS SUMMARY

The nuclear genome was removed from mature MII oocytes, and was fused to an enucleated oocyte of a different donor. Oocytes were activated, allowed to develop to the blastocyst stage and stem cells were derived.

Full Methods and any associated references are available in the online version of the paper.

Received 10 July; accepted 21 November 2012.
Published online 19 December 2012.

1. Jenuth, J. P., Peterson, A. C., Fu, K. & Shoubridge, E. A. Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. *Nature Genet.* **14,** 146–151 (1997).
2. Bolhuis, P. A. *et al.* Rapid shift in genotype of human mitochondrial DNA in a family with Leber's hereditary optic neuropathy. *Biochem. Biophys. Res. Commun.* **170,** 994–997 (1990).
3. Cree, L. M. *et al.* A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nature Genet.* **40,** 249–254 (2008).
4. Wai, T., Teoli, D. & Shoubridge, E. A. The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nature Genet.* **40,** 1484–1488 (2008).
5. Steffann, J. *et al.* Analysis of mtDNA variant segregation during early human embryonic development: a tool for successful NARP preimplantation diagnosis. *J. Med. Genet.* **43,** 244–247 (2006).
6. Nuffield Council on Bioethics. *Novel Techniques for the Prevention of Mitochondrial DNA Disorders: an Ethical Review* http://www.nuffieldbioethics.org/news/discussion-event-novel-techniques-prevention-mitochondrial-dna-disorders-ethical-review (2012).
7. Sato, A. *et al.* Gene therapy for progeny of mito-mice carrying pathogenic mtDNA by nuclear transplantation. *Proc. Natl Acad. Sci. USA* **102,** 16765–16770 (2005).
8. Tachibana, M. *et al.* Mitochondrial gene replacement in primate offspring and embryonic stem cells. *Nature* **461,** 367–372 (2009).

9. Craven, L. *et al.* Pronuclear transfer in human embryos to prevent transmission of mitochondrial DNA disease. *Nature* **465,** 82–85 (2010).
10. Egli, D. *et al.* Reprogramming within hours following nuclear transfer into mouse but not human zygotes. *Nature Comm.* **2,** 488 (2011).
11. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460,** 278–282 (2009).
12. Sathananthan, A. H. *et al.* Centrioles in the beginning of human development. *Proc. Natl Acad. Sci. USA* **88,** 4806–4810 (1991).
13. Kaufman, M. H., Robertson, E. J., Handyside, A. H. & Evans, M. J. Establishment of pluripotent cell lines from haploid mouse embryos. *J. Embryol. Exp. Morphol.* **73,** 249–261 (1983).
14. Draper, J. S. *et al.* Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nature Biotechnol.* **22,** 53–54 (2004).
15. Kim, K. *et al.* Histocompatible embryonic stem cells by parthenogenesis. *Science* **315,** 482–486 (2007).
16. Noggle, S. *et al.* Human oocytes reprogram somatic cells to a pluripotent state. *Nature* **478,** 70–75 (2011).
17. Hyun, C. S. *et al.* Optimal ICSI timing after the first polar body extrusion in *in vitro* matured human oocytes. *Hum. Reprod.* **22,** 1991–1995 (2007).
18. Brinkley, B. R. & Cartwright, J. Jr. Cold-labile and cold-stable microtubules in the mitotic spindle of mammalian cells. *Ann. NY Acad. Sci.* **253,** 428–439 (1975).
19. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).
20. Mimaki, M. *et al.* Reversible infantile respiratory chain deficiency: a clinical and molecular study. *Ann. Neurol.* **68,** 845–854 (2010).
21. Wang, J., Venegas, V., Li, F. & Wong, L.-J. Analysis of mitochondrial DNA point mutation heteroplasmy by ARMS quantitative PCR. *Curr. Protocols Human Genet.* **Chapter 19,** Unit–19.16 (2011).
22. Bai, R.-K. & Wong, L.-J. C. Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach. *Clinical Chem.* **50,** 996–1001 (2004).
23. Lee, H. S. *et al.* Rapid mitochondrial DNA segregation in primate preimplantation embryos precedes somatic and germline bottleneck. *Cell Rep.* **1,** 506–515 (2012).
24. Meirelles, F. V. & Smith, L. C. Mitochondrial genotype segregation during preimplantation development in mouse heteroplasmic embryos. *Genetics* **148,** 877–883 (1998).
25. Jenuth, J. P., Peterson, A. C. & Shoubridge, E. A. Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. *Nature Genet.* **16,** 93–95 (1997).
26. Sharpley, M. S. *et al.* Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* **151,** 333–343 (2012).
27. Fujikura, J. *et al.* Induced pluripotent stem cells generated from diabetic patients with mitochondrial DNA A3243G mutation. *Diabetologia* **55,** 1689–1698 (2012).
28. Birket, M. J. *et al.* A reduction in ATP demand and mitochondrial activity with neural differentiation of human embryonic stem cells. *J. Cell Sci.* **124,** 348–358 (2011).
29. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *New Engl. J. Med.* **350,** 1353–1356 (2004).
30. Inoue, S., Fuseler, J., Salmon, E. D. & Ellis, G. W. Functional organization of mitotic microtubules. Physical chemistry of the *in vivo* equilibrium system. *Biophys. J.* **15,** 725–744 (1975).
31. Bianchi, V., Coticchio, G., Fava, L., Flamigni, C. & Borini, A. Meiotic spindle imaging in human oocytes frozen with a slow freezing procedure involving high sucrose concentration. *Human Reprod.* **20,** 1078–1083 (2005).
32. Larman, M. G., Minasi, M. G., Rienzi, L. & Gardner, D. K. Maintenance of the meiotic spindle during vitrification in human and mouse oocytes. *Reprod. Biomed. Online* **15,** 692–700 (2007).
33. Forman, E. J. *et al.* Oocyte vitrification does not increase the risk of embryonic aneuploidy or diminish the implantation potential of blastocysts created after intracytoplasmic sperm injection: a novel, paired randomized controlled trial using DNA fingerprinting. *Fertil. Steril.* **98,** 644–649 (2012).
34. Gook, D. A., Osborn, S. M., Bourne, H. & Johnston, W. I. Fertilization of human oocytes following cryopreservation; normal karyotypes and absence of stray chromosomes. *Hum. Reprod.* **9,** 684–691 (1994).
35. ASRM. Mature oocyte cryopreservation: a guideline. *Fertil. Steril.* http://dx.doi.org/10.1016/j.fertnstert.2012.09.028 (18 October 2012).
36. Tachibana, M. *et al.* Towards germline gene therapy of inherited mitochondrial diseases. *Nature* http://dx.doi.org/10.1038/nature11647 (this issue).
37. Tachibana, M., Sparman, M. & Mitalipov, S. Chromosome transfer in mature oocytes. *Nature Protoc.* **5,** 1138–1147 (2010).
38. Liu, D., Vader, G., Vromans, M. J., Lampson, M. A. & Lens, S. M. Sensing chromosome bi-orientation by spatial separation of aurora B kinase from kinetochore substrates. *Science* **323,** 1350–1353 (2009).
39. A mother's gift, minus mitochondria. *Nature Med.* **16,** 645 (2010).
40. Society for Assisted Reproductive Technology, American Society for Reproductive Medicine. Assisted reproductive technology in the United States: 2000 results generated from the American Society for Reproductive Medicine/Society for Assisted Reproductive Technology Registry. *Fertil. Steril.* **81,** 1207–1220 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Chang and K. Eggan for discussions, Z. Hall for critical reading of the manuscript, and L. Yu and O. Nahum for SNP-array preparation. We thank anonymous oocyte donors for participating in research, and M. Spencer for a Lykos laser system. This work was supported by the New York Stem Cell Foundation, the New York State Stem Cell Science award C026184, and the Bernard and Anne Spitzer Fund.

Author Contributions M.V.S. consented oocyte donors and retrieved oocytes. R.P. contributed IVF developmental data. R.S.G. and M.V.S. wrote institutional review board and consent documents. D.E., D.P. and S.N. designed and performed experiments with oocytes. D.P. and V.E. determined heteroplasmy. N.T. performed array analysis of single cells. D.E., D.P., V.E., L.S., K.A.W., H.H., M.Z. and D.J.K characterized stem-cell lines. D.E., D.P., V.E. and M.H. wrote the paper.

Author Information Illumina array data have been deposited at the Gene Expression Omnibus (GEO) under accession number GSE42077; Affymetrix array data have been deposited at the GEO under accession number GSE42271. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.E. (d.egli@nyscf.org) or M.S. (mvs9@columbia.edu).

METHODS

Research subjects. Human oocytes were donated by women enrolled in the reproductive egg donation program at Columbia University Medical Center (CUMC). Oocyte donors were given the option to donate for research only after they had enrolled and qualified as reproductive egg donors. Their decision to donate was therefore independent of the decision to donate for research. Compensation was equivalent to that given for reproductive donation, and did not depend on the number or quality of oocytes donated, in agreement with American Society for Reproductive Medicine (ASRM) guidelines[41]. The menstrual cycles of the subjects were synchronized using oral contraceptive pills. On the third day after the discontinuance of the oral contraceptive pills, ovarian hyperstimulation was initiated regardless of menstrual flow, using parenterally administered gonadotropins. Once lead follicles reached 18–22 mm diameter, 4 mg leuprolide acetate was administered to trigger final maturation and oocytes were retrieved 35 h later. With each donation, a skin biopsy and 3–4 ml of blood were taken. All human subject protocols were reviewed and approved by the CUMC institutional review board and the embryonic stem-cell research oversight committee of CUMC. All oocyte donors gave informed consent.

Manipulation of human oocytes. Oocytes were transported in vials containing GMOPSplus (Vitrolife), using a portable incubator heated to 37 °C. A total of 73 MII oocytes was used for this study. Oocytes of different donors were placed in separate drops of GMOPSplus containing 5 μg ml^{-1} cytochalasin B (Sigma) covered with mineral oil (Irvine Scientific). Karyoplasts were aspirated into a pipette with a diameter of 20 μm (Humagen), after incubation in medium containing cytochalasin B for 3–5 min. If the karyoplast contained a larger amount of cytoplasm, the extra cytoplasm was removed by pressing the cytoplasm against the zona pellucida. Karyoplasts of donor 1 were inserted below the zona pellucida of an enucleated oocyte of donor 2, and fused using either inactivated Sendai virus HVJ-E (GenomeOne, Cosmo Bio), diluted with fusion buffer 1:40, or electrofusion, performed in cell fusion medium (0.26 M mannitol, 0.1 mM $MgSO_4$, 0.05% BSA and 0.5 mM HEPES) using 2–8 fusion pulses of 20 μs width and 1.3 V cm^{-1} strength (LF201, NEPA Gene). Each exchange required approximately 10 min. After aspiration of no more than two karyoplasts, transfer was undertaken, unless karyoplasts were used for cryopreservation. All manipulations were performed on a 37 °C heated stage (Tokai Hit) of a Nikon TE 2000U inverted microscope, using Narishige micromanipulators. The transfer of the oocytes from the incubator to the inverted microscope required their brief exposure to room temperature (21 °C). Oocyte culture was conducted in Global medium supplemented with 10% plasmanate (Talecris), or in Global total (LifeGlobal) in a Minc incubator (Cook Medical), infused with a defined gas mixture of 6% CO_2, 5% O_2 and 89% N_2. Parthenogenetic activation of oocytes was done using ionomycin, followed by incubation in 10 μM puromycin for 4 h. Manipulation of oocytes was completed within 3–5 h after aspiration.

Vitrification and thawing of all oocytes was achieved through the use of the cryotop kit (Kitazato) and used in accordance with the manufacturer's instructions on an unheated stage of a Nikon SMZ1500. For vitrification, oocytes were placed into basic solution with equilibrium solution added over a 15-min period. Oocytes were transferred to vitrification solution and then immediately placed onto the cryotop, with minimal solution carry over, and plunged directly into liquid nitrogen. The cryotop was placed into a protective straw for storage. After thawing, the cryotop was quickly placed into pre-warmed (37 °C) thawing solution for 1 min. Oocytes were then transferred into diluent solution (21°C) and washing solution (21 °C) over a 9-min period, after which they were transferred to regular culture medium or GMOPSplus. Ten out of eleven cryopreserved oocytes were viable after thawing. Statistical analysis of development and spontaneous activation was undertaken using the chi-square test. $P < 0.05$ was considered statistically significant.

Stem-cell derivation and analysis. Parthenotes were allowed to develop to the expanded blastocyst stage, or day 6 or 7 after activation. The trophectoderm of blastocysts were ablated using the Lykos Laser (Hamilton Thorne), as previously described[42]. Isolated ICMs were plated on mouse embryonic fibroblast feeder (MEF) layers in stem-cell culture medium (hESm; KO-DMEM with high glucose, supplemented with 20% knockout serum replacement and bFGF; all cell culture reagents from Life Technologies). For pluripotency analysis, stem cells were fixed and stained for OCT4 (also known as POU5F1), SOX2 (both Stemgent), NANOG (Cell Signaling Technology), SSEA3, SSEA4 (both R&D Systems), TRA-1-60 and TRA-1-80 (Millipore). Images were taken using an Olympus IX71 epifluorescence microscope. Live cultures were sent to Cell Line Genetics or WiCell for karyotyping and STR genotyping. Gene expression analysis was undertaken using the Illumina HumanHT-12 Expression BeadChip and analysed using the Illumina Beadstudio software. Quantitative gene expression analysis was undertaken using the Nanostring nCounter system and analysed using nSolver (Nanostring Technologies). Teratomas were generated by subcutaneous injection into NSG mice (Jackson Laboratories) and collected after 10–15 weeks. Animal experimentation was performed under a Columbia Institutional Animal Care and Use Committee protocol.

Nuclear and mitochondrial genotyping. Samples for both nuclear and mitochondrial genotyping were prepared using the high-pure template preparation kit as per the manufacturer's instructions (Roche). Using a range of nuclear DNA and mtDNA primers (Supplementary Table 8), PCR products were generated using Red-Taq (Sigma) or Blue-Taq (Denville Scientific) as per the manufacturer's instructions and purified using the high-pure PCR product purification kit (Roche). Sanger sequencing was undertaken via either Genewiz or Macrogen. Sequences were analysed using ApE (http://biologylabs.utah.edu/jorgensen/wayned/ape/). Complete mitochondrial genotyping was undertaken using the Affymetrix GeneChip human mitochondrial resequencing array 2.0 (MitoChip v.2.0), according to the manufacturer's recommended protocols (primers and condition for the long PCR described in Supplementary Table 8). Sequences were directly compared to the revised Cambridge Reference Sequence for human mtDNA (NCBI accession NC_012920). Genotyping of nuclear DNA was done using Affymetrix GeneChip human mapping 250 K NspI arrays according to manufacturer's instructions. Analysis was performed using Affymetrix genotyping console.

Copy number variation analysis. Copy number analysis was performed using Affymetrix 6.0 SNP arrays as per the manufacturer's instructions. CNVs were detected using NEXUS 6 and the SNP-FASST2 segmentation algorithm. High gains were set to the threshold of 0.7, gains at 0.1, losses at -0.15 and big losses at -1.1. A significance threshold of 5.0×10^{-7} was used, with minimum number of 10 probes required per CNV call and a minimum size of 50 kilobases (kb). Copy losses/gains were analysed to a minimum of 10 kb. After participation analysis to determine whether CNVs were new or preexisting, each call was manually inspected for visual confirmation of the call. Either a *t*-test or a one-way analysis of variance (ANOVA) with Bonferroni's multiple comparison test was used for statistical analysis. $P < 0.05$ was considered to be statistically significant.

Quantification of mtDNA copy number. Quantitative real-time PCR analysis of mtDNA copy number was achieved using previously designed primers and calculating unknown samples on a standard curve plotting copy number against a mean threshold value. A standard curve was generated using serial dilutions of a purified-PCR product generated using Red-Taq (Sigma) and previously published primers designed for the nucleotide positions 8259–8273 and 8475–8489 (ref. 43; Supplementary Table 8). Quantification of mtDNA copy number was achieved through the use of primers designed for the nucleotide positions 8290–8308 and 8418–8438. Samples were prepared using the high-pure template purification kit as per the manufacturer's instructions, although samples were eluted in only 30 μl of elution buffer (Roche). Reaction mixtures were prepared (in triplicate) with 3 μl template DNA, 5 μl 2× SYBR Green PCR master mix (Promega), 100 nM of each primer and water to a final volume of 10 μl. The reactions were performed in a Stratagene MX3000P with the following cycle: hold at 95 °C for 10 min followed by 40 cycles of 95 °C for 15 s, 53 °C for 30 s, and 72 °C for 1 min. The copy number of an unknown sample was calculated from the standard curve and adjusted for the dilution factor.

Estimation of oocyte and karyoplast volume. Oocytes and karyoplasts were imaged on either an Olympus IX71 or a Zeiss LSM5 PASCAL microscope following staining with MitoTracker Red (Life Technologies) and either Hoescht (Sigma) or Draq5 (Biostatus). Single images were taken at the midpoint of each sample of interest, with complete *Z* stacks also imaged. Images were analysed using Zen LE (Zeiss) and based on the measured diameter, the volume was calculated (with the assumption that oocytes and karyoplasts are spherical).

Quantification of heteroplasmy levels: ARMS-qPCR and last-hot cycle RFLP. Heteroplasmy was analysed using both last-hot cycle RFLP and ARMS-qPCR. The presence of three SNPs, m.1670A>T (*MT-TV*), m.4715A>G (*MT-ND2*) and m.16129A>G (non-coding region), was validated by Sanger sequencing and the polymorphisms were analysed by RFLP. The regions flanking the variations site were PCR-amplified (primers and PCR condition described in Supplementary Table 8). The 305-base-pair (bp) fragment containing the m.4715A variant was digested by BspEI into two fragments (204 and 101 bp), whereas the m.4715G sequence lacked the BspEI recognition site (Supplementary Fig. 7a). The 387-bp fragment containing the m.16129A variant was digested by BanI into two fragments (176 and 211 bp), whereas the m.16129G mtDNA had an additional BanI recognition site yielding three fragments (176, 130 and 81 bp) (Supplementary Fig. 7b). The 312-bp fragment containing the m.1670T variant was digested by AluI in two fragments (120 and 192 bp), whereas the m.1670A had an additional AluI recognition site yielding to three fragments (120, 60 and 132 bp) (Supplementary Fig. 7c). To assess heteroplasmy, [α-^{32}P]deoxycytidine 5′-triphosphate (dCTP; 3,000 Ci mmol^{-1}) (Perkin Elmer Health Science) was added to the last PCR cycle, the hot-labelled digested products were electrophoresed in a

10% non-denaturing acrylamide gel, and the bands analysed in a PhosphorImager (Typhoon TRIO variable mode imager, GE Healthcare Life Science) using ImageQuant TL v.7.01 software (GE Healthcare). The experiments were performed in duplicates using oocyte donor fibroblasts (1110 and 1111), swaPS1 and swaPS2 cells (P2, P9 and P40), cells differentiated in pancreatic β-cells, and swaPS3 cells (P4, P5, P6, P8, P10, P13, P14 and P20).

ARMS-qPCR primers were designed to amplify specifically the DNA of only one donor with two homozygous SNPS used at either end of the sequence based on the sequencing of HVR1 (Supplementary Table 6). Furthermore, the primers were designed such that the product generated would contain internal SNPs that could be verified by Sanger sequencing to confirm donor-specific product amplification (Supplementary Fig. 8a). To create a set of standards to confirm accuracy of the assay, the HVR of the various donors was amplified using Red-Taq (Sigma), and the product purified with the high-pure PCR product purification kit (Roche). The concentration of PCR product was calculated using a NanoDrop spectrophotometer (NanoDrop Technologies), and the copy number was calculated based on a standard curve (10^1–10^7 copies) generated with the assumption that DNA has a molecular mass of 650 daltons (Da) (primers and conditions in Supplementary Table 8). From this, standards were generated at various percentages of the two donors to confirm accuracy of the ARMS-qPCR assay. Expected values for the standards were matched against measured values using the equation: Mutant heteroplasmy level (%) $= 1/(1 + (1/2)\Delta C_T) \times 100\%$, in which $\Delta C_T = C_{T\text{wild type}} - C_{T\text{mutant}}$ (Supplementary Fig. 8b–d). After confirmation, reactions were run under conditions d, e or f (as described in Supplementary Table 8), with samples of unknown heteroplasmy run alongside positive and negative controls and calculated using the above equation. Samples were run in triplicate in a reaction containing 0.6 ng μl^{-1} specific donor sample, 500 nM of each primer, 5 μl of 2× SYBR green (Promega) and water to a final volume of 10 μl and analysed using a Stratagene MX300P (Agilent). The determination of homoplasmy was undertaken using primers designed around the HVR of donors 1110 and 1111, allowing the incorporation of five SNPs per primer pair to facilitate primer specificity (Supplementary Table 6). Target DNA was amplified using Blue-Taq (Denville Scientific) under cycle conditions k (as described in Supplementary Table 8), before being analysed on a 2% agarose gel. Subsequently, 1 μl of the PCR product was diluted 1:100 and re-amplified under the same conditions for a further 30 cycles before being analysed on a 2% agarose gel.

High-resolution northern blot analysis. Total RNA from cultured primary fibroblasts grown in a 10 cm^2 dish was extracted using Trizol reagent (Life Technologies) according to the manufacturer's instructions. Large RNA species were precipitated by the addition of 10 mol l^{-1} LiCl, allowing smaller RNAs to be precipitated from the resulting supernatant. Small RNAs (1.5 μg) were denaturated (70 °C for 5 min) and separated through an 8%, 8 mol l^{-1} urea denaturating polyacrylamide (19:1) gel using 0.5× Tris-borate EDTA (TBE) as running buffer. Separated samples were electroblotted onto Nytran SuPerCharge TurboBlotter membrane (Whatman) using a TE77X semi-dry transfer unit (Hoefer) and immobilized by ultraviolet cross-linking. Regions of mtDNA encompassing the $tRNA^{Val}$ and $tRNA^{Leu(UUR)}$ genes were amplified using specific primers (see Supplementary Table 8). Purified PCR products were radiolabelled with [α-^{32}P]dCTP (3,000 Ci $mmol^{-1}$) by random primer method, and unincorporated nucleotides were removed by gel filtration through a Sephadex G-50 DNA grade column (Amersham Pharmacia Biotech). Hybridization was performed at 42 °C overnight using a QuikHyb hybridization solution (Agilent Technologies Stratagene Products Division) containing 500,000 c.p.m. radiolabelled probes. After hybridization, two 15-min washes were performed at room temperature in 2× SSC and 0.1% SDS, followed by one 30-min wash in 2× SSC and 0.1% SDS at 60 °C. Blots were subjected to PhosphorImager analysis and the radioactive signal for the mt-$tRNA^{Val}$ probe (69 bp) normalized to the $tRNA^{Leu(UUR)}$ signal (75 bp).

Directed differentiation of swaPS lines. To confirm the heteroplasmy levels in terminally differentiated tissues, swaPS1 and swaPS2 were differentiated along endodermal, ectodermal and mesodermal lineages. β-cell differentiation was performed as described[44], with the addition of calcium chelator, EGTA (75 μM), on day 1 and an activin receptor-like kinase inhibitor, SB431542 (2 μM), on days 9–12 of differentiation. Staining for SOX17, PDX1 (both R&D Systems) and C-peptide (Millipore) was undertaken at days 3, 10 and 14. DNA samples for heteroplasmy analysis were collected at day 14. The ectodermal differentiation was undertaken following the previously described dual-SMAD protocol[45]. After 2 weeks and two passages as neural progenitors, further differentiation was induced through the addition of BDNF (10 ng ml^{-1}, R&D Systems). After a further 3 weeks, cells were fixed for immunostaining, or DNA was collected for heteroplasmy analysis. Antibodies used for staining included TUJ1 (Sigma), MAP2 (Abcam), nestin, neurofilament (both Millipore) and SOX2 (Stemgent). Cells of the mesodermal lineage were generated using two protocols. First, the generation of contracting cardiomyocytes (from swaPS1, swaPS2 and swaPS4) was achieved using a previously published protocol[46], with videos recorded using an Olympus IX71 with the DP2-BSW software. Second, the differentiation of swaPS1 and swaPS2 into fibroblasts was achieved by growing undifferentiated stem cells in human fibroblast medium (hFm; DMEM supplemented with 10% FBS, 1% penicillin/streptomycin and 1% glutamax) for 2–4 weeks with a single passage during this time. After 14–28 days of growth, cells were sorted by FACS to enrich for TRA-1-60$^-$ SSEA4$^-$ CD56$^-$ CD13$^+$ cells (Supplementary Fig. 9a). Antibodies used for staining included αSMA, phalloidin (both Sigma) and CD-13 (BD Biosciences). Gene expression analysis of the swaPS fibroblasts was undertaken using the Nanostring nCounter system as previously described. After differentiation, swaPS fibroblasts were placed into hESm and grown for 3 weeks on a feeder layer. No colonies were visible during this time period, and flow cytometry confirmed the absence of stem-cell-positive markers indicating swaPS fibroblasts could not spontaneously revert to a stem-cell state (Supplementary Fig. 9b, c).

Generation of iPS cell lines. Biopsies were taken using a biopsy kit (AccuPunch, Accuderm) after local anaesthesia using lidocaine (1%, Hospira). Punch biopsies (3 mm) were cut into 10–15 small pieces from which fibroblasts were allowed to grow for 4 weeks. Fibroblasts were then passaged using TrypLE and plated in hFm at a density of 50,000 cells per well (6-well plate) overnight before being infected with the Cytotune iPS Sendai reprogramming kit as per the manufacturer's instructions (all reagents from Life Technologies). Infected cells were grown in hESm containing three additional compounds[47] (thiazovivin, SB431542 and PD0325901; all Stemgent) for 10 days before FACS enrichment of SSEA4$^+$ Tra-1-60$^+$ CD13$^-$ cells. Colonies were picked 7–14 days later and pluripotency was confirmed through the staining of pluripotency markers as described above and Nanostring gene expression analysis as previously described (Supplementary Fig. 9d–h).

Metabolic analysis. Stem-cell lines were transferred from growing on MEFs to Matrigel (BD Biosciences)-coated plates and cultured in m-TeSR (StemCell Technologies). Cells were grown to approximately 90% confluence in 10-cm^2 dishes before being collected and stored at −80 °C. Biochemical activities of COX, NADH-cytochrome *c* reductase (complex I + III), succinate-cytochrome *c* reductase (complex II + III), NADH-CoQ reductase (complex I), succinate dehydrogenase (complex II) and citrate synthase were assayed spectrophotometrically as previously described[48]. Respiratory chain enzyme activity values were normalized to citrate synthase, an index of mitochondrial mass, and data were expressed as mean ± s.d. of at least two experiments. Biochemical activities were measured in three pES cell lines, two swaPS cell lines, two human ES cell lines and three iPS lines. Metabolic analysis was also performed in swaPS-derived fibroblasts (two cell lines) compared to control fibroblasts (six cell lines). One-way ANOVA with Bonferroni's multiple comparison test was used to compare groups. $P < 0.05$ was considered to be statistically significant. Live metabolic analysis of cells was undertaken using the Seahorse stress-test kit as per the manufacturer's instructions. In brief, 42,500 cells were seeded into the assay plate and allowed to grow overnight. The next day cells growth media was replaced with XF assay medium (Seahorse) supplemented with 25 mM glucose, 0.4% BSA (both Sigma) 1% glutamax and 1% sodium pyruvate (both Invitrogen) for 1 h. After 1 h, cells were analysed using the XF24 (Seahorse). Experiments were performed in duplicates with an *n* of 4 per group. Results were analysed using the Seahorse XF24 software one-way ANOVA with Bonferroni's multiple comparison test was used to compare groups. $P < 0.05$ was considered to be statistically significant.

41. The Ethics Committee of the American Society for Reproductive Medicine. Financial compensation of oocyte donors. *Fertil. Steril.* **88,** 305–309 (2007).
42. Chen, A. E. *et al.* Optimal timing of inner cell mass isolation increases the efficiency of human embryonic stem cell derivation and allows generation of sibling cell lines. *Cell Stem Cell* **4,** 103–106 (2009).
43. Lin, D. P.-C. *et al.* Comparison of mitochondrial DNA contents in human embryos with good or poor morphology at the 8-cell stage. *Fertil. Steril.* **81,** 73–79 (2004).
44. D'Amour, K. A. *et al.* Production of pancreatic hormone–expressing endocrine cells from human embryonic stem cells. *Nature Biotechnol.* **24,** 1392–1401 (2006).
45. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnol.* **27,** 275–280 (2009).
46. Burridge, P. W. *et al.* A universal system for highly efficient cardiac differentiation of human induced pluripotent stem cells that eliminates interline variability. *PloS ONE* **6,** e18293 (2011).
47. Lin, T. *et al.* A chemical platform for improved induction of human iPSCs. *Nature Methods* **6,** 805–808 (2009).
48. DiMauro, S. *et al.* Cytochrome *c* oxidase deficiency in Leigh syndrome. *Ann. Neurol.* **22,** 498–506 (1987).

doi:10.1038/nature11843

Crystal structure of Prp8 reveals active site cavity of the spliceosome

Wojciech P. Galej[1], Chris Oubridge[1], Andrew J. Newman[1] & Kiyoshi Nagai[1]

The active centre of the spliceosome consists of an intricate network formed by U5, U2 and U6 small nuclear RNAs, and a pre-messenger-RNA substrate. Prp8, a component of the U5 small nuclear ribonucleoprotein particle, crosslinks extensively with this RNA catalytic core. Here we present the crystal structure of yeast Prp8 (residues 885–2413) in complex with Aar2, a U5 small nuclear ribonucleoprotein particle assembly factor. The structure reveals tightly associated domains of Prp8 resembling a bacterial group II intron reverse transcriptase and a type II restriction endonuclease. Suppressors of splice-site mutations, and an intron branch-point crosslink, map to a large cavity formed by the reverse transcriptase thumb, and the endonuclease-like and RNaseH-like domains. This cavity is large enough to accommodate the catalytic core of group II intron RNA. The structure provides crucial insights into the architecture of the spliceosome active site, and reinforces the notion that nuclear pre-mRNA splicing and group II intron splicing have a common origin.

Removal of introns from nuclear pre-mRNA occurs in two consecutive trans-esterification reactions catalysed by a multi-megadalton, dynamic RNA–protein complex known as the spliceosome (reviewed in ref. 1). The spliceosome is formed on pre-mRNA substrates from its five canonical subunits, the small nuclear ribonucleoprotein particles (U1, U2, U4/U6 and U5 snRNPs), and numerous non-snRNP factors. Assembly of the spliceosome begins with the recognition of the 5′-splice site (5′-SS) by U1 snRNP, followed by recognition of the sequence flanking a specific adenosine in the intron, termed the branch point, by U2 snRNP. After the recruitment of a pre-assembled U5–U4/U6 tri-snRNP, in which U4 and U6 small nuclear RNAs (snRNAs) are extensively base-paired, the spliceosome undergoes a major structural and compositional rearrangement including unwinding of the U4/U6 snRNA duplex and concomitant formation of a highly structured RNA network between U2, U5 and U6 snRNAs and the 5′-SS and branch-point sequences in the pre-mRNA. This leads to a nucleophilic attack of the 2′-OH of the branch-point adenosine at the 5′-SS, producing exon 1 and lariat intron–exon 2 intermediates. Further remodelling enables a nucleophilic attack of exon 1 at the 3′-splice site (3′-SS), yielding spliced mRNA and lariat intron products. At the catalytic core of the spliceosome, the base-paired U2–U6 snRNAs provide a platform for correct positioning of the branch point and 5′-SS (refs 2–6) as well as coordinating catalytic magnesium ions[7,8], and U5 snRNA aligns the exons for the second catalytic step[9–11].

Three U5 snRNP proteins, Prp8, Brr2 and Snu114, have crucial roles in the activation of the spliceosome and the formation of the catalytic core for the two trans-esterification reactions. Yeast Prp8 is a 280-kilodalton (kDa) protein and has 61% sequence identity to its human counterpart. Human PRP8 (encoded by *PRPF8*) forms a salt-stable complex with the EF2-like GTPase SNU114 (encoded by *EFTUD2*) and the DExD/H-box family helicase BRR2 (also known as SNRNP200)[12]. In yeast, GTP-bound Snu114 activates Brr2 (refs 13, 14), which unwinds the U4/U6 snRNA duplex to allow U6 snRNA to base-pair with U2 snRNA. Prp8 crosslinks with crucial snRNAs (U5 and U6) and substrate residues (5′-SS, 3′-SS and branch point)[15–20]. Many Prp8 mutations suppress splicing defects caused by mutations in the 5′-SS, 3′-SS and branch point (ref. 21 and references therein). Hence, Prp8 is located at the heart of the spliceosome. Despite the central role of Prp8 in splicing, the structures of only two small domains, the RNaseH-like and Jab1/MPN domains, have been determined so far[22–26]. On the basis of bioinformatic analysis, it has been proposed that part of Prp8 forms an RNA recognition motif (RRM)[21], and sequence similarity was recently reported[27] between the central part of Prp8 and a catalytic domain of reverse transcriptase from bacterial group II intron-encoded protein (IEP).

In the cytoplasm, Prp8 forms a large complex containing U5 snRNA, Snu114, seven Sm proteins (B, D1, D2, D3, E, F and G) and the U5 snRNP assembly factor Aar2. After import of this complex into the nucleus, Aar2 is replaced by Brr2, and other proteins are recruited to form the mature U5 snRNP[28,29]. Here we report a crystal structure of a large carboxy-terminal fragment of *Saccharomyces cerevisiae* Prp8 (residues 885–2413) in complex with full-length Aar2.

Overall architecture of the complex

A 176-kDa fragment of yeast Prp8 was co-crystallized with Aar2. The crystals diffracted to 1.9 Å resolution and the structure was solved by molecular replacement using the crystal structures of the RNaseH-like and Jab1/MPN domains and Aar2 (Protein Data Bank (PDB) accessions 3SBT and 2OG4) as search models (Methods and Supplementary Tables 1 and 2). A methylmercury derivative was used to verify molecular replacement solutions (Supplementary Fig. 1). The structure was refined at 2 Å resolution to an R_{free} value of 24.8% (Supplementary Fig. 2). The crystal structure revealed a new large domain of Prp8 (residues 885–1824) spanning the entire length of the complex (Figs 1a and 2a). The RNaseH-like[22–24] and Jab1/MPN[25,26] domains, each connected by disordered linkers, fold back to interact with the new domain via Aar2 (Fig. 2 and Supplementary Fig. 3). This domain can be subdivided into a large polymerase-like domain (Supplementary Table 3) and a small type II restriction endonuclease-like domain (Supplementary Table 4) that interact intimately through the linker domain. The polymerase-like domain is composed of three canonical subdomains[30]: palm, fingers and thumb (Supplementary Figs 4 and 5). They form a deep cleft, which accommodates the nucleic

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK.

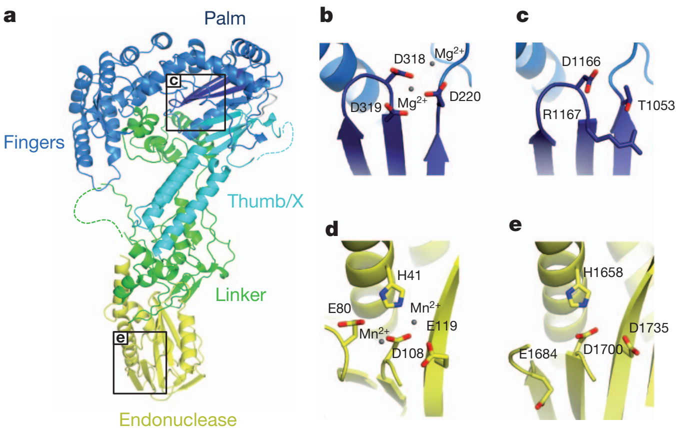


Figure 1 | Structure of the large domain in yeast Prp8 (residues 885–1824). **a**, The large domain consists of a group II intron reverse transcriptase-like domain and a type II restriction endonuclease-like domain. **b**, The palm subdomain of hepatitis C virus RNA-dependent RNA polymerase (1NB6). Asp residues (Asp 220 in motif A, and Asp 318 and Asp 319 motif C) coordinate two catalytic Mg^{2+} ions. **c**, The corresponding residues in the palm subdomain of the group II intron reverse transcriptase-like domain of Prp8. **d**, The catalytic centre of the endonuclease domain of the influenza virus polymerase acidic subunit (2W69). His 41, Glu 80, Asp 108 and Glu 119 coordinate two catalytic divalent ions. **e**, The corresponding residues in the endonuclease domain of Prp8.

acid template and primer in polymerases. In all polymerases, the most highly conserved palm subdomain lies at the bottom of the deep cleft and contains the catalytically important residues embedded in four conserved motifs (A, B, C and D)[30]. Aspartates, one in motif A and two in motif C, coordinate two Mg^{2+} ions, required for catalysis (Fig. 1b), whereas motif B is involved in nucleotide selection. In Prp8, the palm subdomain forms a four-stranded anti-parallel β-sheet (RTβ4 and RTβ7–RTβ9) flanked by α-helices (RTα6 and RTα9–RTα13). Motif C is located in the loop between RTβ7 and RTβ8, whereas motifs A, B and D are in RTβ4, RTα9 and RTα13, respectively. Only one of the three aspartate residues (Asp 1166), equivalent to the first aspartate in the Tyr-X-Asp-Asp consensus sequence in motif C[30], is conserved in Prp8 (Fig. 1c and Supplementary Fig. 3). Thr 1053 and Arg 1167, which replace the two other catalytic aspartates, are neither capable of metal ion coordination, nor conserved in Prp8 from different species. Hence, the 'active site' of the Prp8 polymerase-like domain is unlikely to bind divalent metal ions. Residues 1048–1182 of yeast Prp8 were recently reported[27] to show sequence similarity to the region corresponding to the reverse transcriptase palm domain of bacterial group II IEP. Hence, we will more appropriately refer to this domain as the reverse transcriptase domain. Residues 1058–1151 were predicted to form an RRM[21] but this region does not resemble the RRM fold, and is embedded in the finger-like subdomain (Supplementary Fig. 6). The thumb subdomain (1257–1375) forms a four-stranded anti-parallel β-sheet (RTβ10–RTβ13), followed by a three-helix bundle (RTα14–RTα16). A significant sequence similarity between this region and the thumb/maturase X (Th/X) domain of fungal group II intron reverse transcriptase was noted and it was correctly predicted to form a helical bundle[27] (Supplementary Fig. 7).

Residues 1650–1810 adopt a type II restriction endonuclease-like fold, with characteristic five mixed β-strands (Enβ1, Enβ2 and Enβ4–Enβ6) flanked by three α-helices (Enα1–Enα3). The Prp8 endonuclease domain is structurally most similar to the endonuclease domain of the influenza virus polymerase acidic (PA) subunit[31,32], with little apparent sequence conservation (9% identity) (Supplementary Fig. 8 and Supplementary Table 4). Two glutamates (Glu 80 and Glu 119), one aspartate (Asp 108) and one histidine (His 41) are involved in

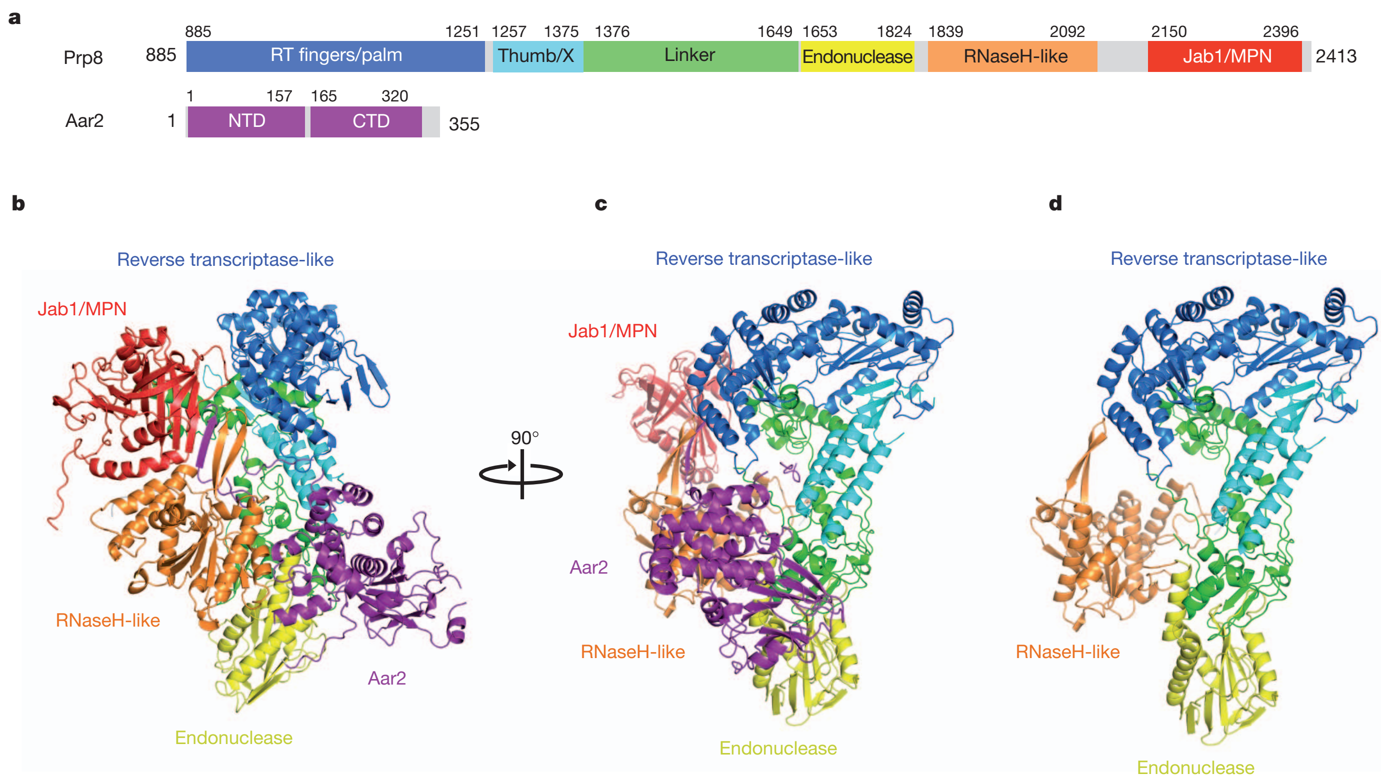


Figure 2 | Overall structure of yeast Prp8$^{885-2413}$ in complex with Aar2. **a**, Domain architecture of Prp8$^{885-2413}$ and Aar2. CTD, C-terminal domain; NTD, N-terminal domain; RT, reverse transcriptase. **b**, Aar2 organizes the arrangement of the RT/En, RNaseH-like and Jab1/MPN domains. **c**, Orthogonal view of the complex. **d**, A view (as in **c**) without Aar2 and the Jab1/MPN domain. The RNaseH-like domain has no direct contact with the RT/En domain.

two-metal ion coordination essential for catalytic activity in the PA endonuclease domain[31,32]. Intriguingly, although all of these residues (Fig. 1d, e) are highly conserved in Prp8 except Glu 1684 (Supplementary Fig. 3), replacement of these residues with uncharged amino acids (Asp to Asn, and Glu to Gln), individually and in combination, had no effect on viability (Supplementary Fig. 9). These residues form a network of polar interactions and stabilize the polypeptide loops that block the active site (Supplementary Fig. 10), and hence may be conserved for a structural reason.

Domain organization

The reverse transcriptase/endonuclease (RT/En), RNaseH-like and Jab1/MPN domains are connected by disordered linkers but form a large assembly stabilized by a network of pivotal interactions involving Aar2 (Fig. 2b); this domain arrangement is undoubtedly crucial in the biogenesis of U5 snRNP[28,29]. Aar2 interacts with Prp8 across the junction between the linker and endonuclease domains burying 1,230 $Å^2$ of solvent-accessible surface (Supplementary Fig. 11). The C-terminal tail (Gly 318 to Pro 355) of Aar2 extends from its main body through the cleft between the reverse transcriptase fingers and Th/X domains; its very C-terminal end (residues 348–353) forms a remarkable intermolecular, parallel β-sheet, zipping together the β-hairpin of the RNaseH-like domain (residues 1860–1864) and the β-barrel of the Jab1/MPN domain (2167–2171) (Fig. 2b and Supplementary Fig. 12). This accounts for the crucial role of the C-terminal tail of Aar2 (residues 331–354) in bringing the RNaseH-like and Jab1/MPN domains together[29]. Furthermore, the C-terminal helical domain (α5 and α6) of Aar2 interacts with RHα1 and RHα2 of the RNaseH-like domain burying 433 $Å^2$ of solvent-accessible surface (Fig. 2b and Supplementary Fig. 13). The previously reported interface between Aar2 and the RNaseH-like domain[29] was not observed in our structure. However, one of the crystal contacts present in the Aar2–RNaseH-like domain complex structure (3SBT) is remarkably similar to the interface between the helical regions observed in our complex (Supplementary Fig. 14). This raises the possibility that the biological interface was incorrectly assigned previously. The Jab1/MPN domain makes contact with the reverse transcriptase and linker domains (Supplementary Fig. 15). Individually, these interactions may not be sufficient to fix the Jab1/MPN domain in position, but the network of interactions between the four domains (RT/En, RNaseH-like, Jab1/MPN and Aar2) holds them together. The RNaseH-like domain itself has little if any direct contact with the RT/En domain and is positioned relative to the RT/En domain indirectly by Aar2 (Fig. 2c). Hence, the exact position of the RNaseH-like and Jab1/MPN domains with respect to the RT/En may be altered when Aar2 is replaced by Brr2 and Prp8 forms a complex with Snu114 and Brr2. Indeed, comparison of the $P2_12_12_1$ and $C222_1$ crystal forms revealed movements of the RNaseH-like and Jab1/MPN domains with respect to the RT/En domain (Supplementary Fig. 16). The domain interactions could be modulated when Snu114 binds different nucleotides.

The active site of the spliceosome

The catalytic centre of the spliceosome includes an intricate network of interactions involving U2, U5 and U6 snRNAs and substrate pre-mRNA (Supplementary Fig. 17). Prp8 crosslinks to crucial residues in U6 snRNA and in the invariant exon-binding loop 1 of U5 snRNA as well as to all three sites of chemistry in the pre-mRNA (5′-SS, branch point and 3′-SS)[15–18,33,34]. Contacts between yeast Prp8 and catalytic core RNA residues were previously mapped by crosslinking and proteolytic cleavage; almost all of the crosslinks lie within the RT/En domain of Prp8 (ref. 20; Supplementary Fig. 17). We have used substrates with modified 3′-SS and captured spliceosomes by the nineteen complex (NTC) protein Prp19 or the step 2 factor Prp18 to focus on crosslinks made just before catalytic step 2 (C. M. Norman and A.J.N., unpublished observations). Crosslinks between Prp8 and the residue two nucleotides downstream of the branch point (BP+2) were mapped to the region between residues 1585 and 1598. This disordered region is located between the blue spheres in Fig. 3c. The crosslinking site is located in the mobile loop near the reverse transcriptase Th/X domain and is distant from the residues corresponding to the Mg^{2+}-coordinating residues in the RT/En domains.

Splice site and branch-point suppressors

Screening for suppression of splicing defects caused by mutations in the 5′-SS, 3′-SS and branch point led to the isolation of many Prp8 suppressors[35,36] (see references in ref. 20 and Supplementary Table 5). Most suppressor mutations are located on the concave surface of the Th/X and endonuclease domains facing the RNaseH-like domain (Fig. 3). This space is lined with extended polypeptide chains making few contacts. These loops are part of the regions crosslinked to a nucleotide immediately upstream of the 5′-SS (5′-SS−1), and to BP+2, U6 snRNA (U54) and the U5 snRNA loop 1 (U97)[20]. The surface of the RNaseH-like domain facing this cavity includes Glu 1960 and Glu 1834, which are sites of 5′-SS and 3′-SS suppressor mutations[36]; they are positioned close to the 5′-SS–hPRP8 crosslink[18] (Fig. 3). These site-specific crosslinks together with the suppressor

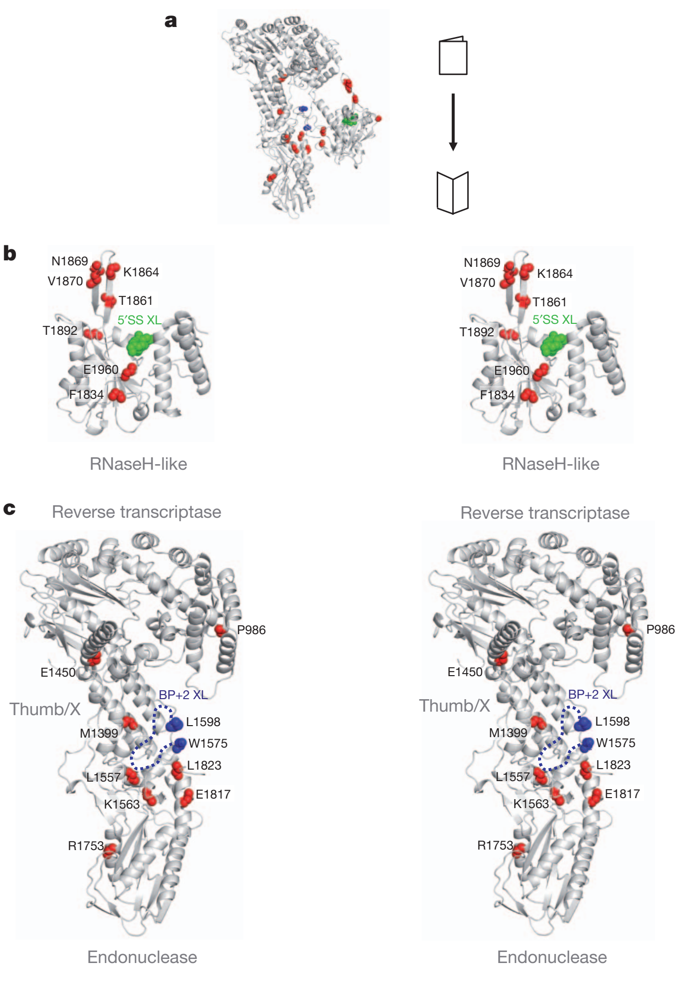


Figure 3 | Overview of the Prp8 active site cavity in an 'open book' view. **a**, Overview with the suppressors of splice site (5′-SS, 3′-SS and branch point) mutations (red spheres). Green spheres indicate the sequence (1966-Ser-Ala-Ala-Met-Ser-1970) corresponding to the crosslinking site of hPRP8 to the 5′-SS (ref. 18). **b**, Stereo view of the RNaseH-like domain surface making up the active site cavity. **c**, Stereo view of the RT/En domain surface making up the active site cavity. Crosslink of the pre-mRNA branch point (BP+2) nucleotide is located between residues 1585 and 1598 in sequence (C. M. Norman and A.J.N., unpublished observations). This site is found within the disordered loop (blue dotted line) between residues 1575 and 1598 (blue spheres).

mutations unambiguously locate the active site of the spliceosome to the cavity formed by the thumb subdomain/endonuclease (Th/En) and RNaseH-like domains.

Four suppressors of 5′-SS and 3′-SS and branch-point mutations map in the RNaseH-like domain β-finger. As in the Aar2–Prp8 complex, the β-finger may be involved in alternative interactions to position the RNaseH-like domain in the Brr2-containing mature U5 snRNP, and these mutations may affect the interaction between Prp8 and the RNA network as previously proposed[24]. Some of the suppressors facilitate the first step but inhibit the second (first-step alleles), and others do the opposite[35] (Supplementary Table 5). It has been proposed that Prp8 has two alternative states that facilitate either the first or the second step[24,35]. It is plausible that this corresponds to alternative positions of the RNaseH-like domain with respect to the RT/En domain; the RNaseH-like domain may in turn transmit conformational changes of the RNA network after each trans-esterification reaction. Comparison of the two crystal forms confirmed flexibility of the RNaseH-like domain, revealing its considerable interdomain movements with respect to reverse transcriptase and endonuclease domains (Supplementary Fig. 16).

Prp8 and spliceosome activation

In the U4-cs1 (cold-sensitive) mutant the three nucleotides (AAA) adjacent to helix I of the U4/U6 snRNA duplex are replaced by UUG, extending helix I by three base-pairs and concealing part of the U6 snRNA sequence (ACAGA) that base-pairs with the 5′-SS. At the restrictive temperature (16 °C) the spliceosome stalls before the first trans-esterification because the U4/U6 snRNAs remain base-paired and the ACAGA box fails to base-pair with the 5′-SS. Screens for suppressors of U4-cs1 in Prp8 isolated many single mutations in five regions (Prp8-cat mutants; regions a–e)[36,37] (Supplementary Table 6). Two of the five regions (d and e) are within our crystal structure (Fig. 4). Six mutations in region d map within or near the four-stranded β-sheet protruding from the reverse transcriptase fingers domain (Fig. 4c), and three map on the loop connecting RTα12 and RTα13. These are the most exposed parts of the fingers and palm domain. In region d and part of region e, three of the mutations are located within or near the exposed β-hairpin of the reverse transcriptase-like domain and two on the exposed loop in the endonuclease domain. Intriguingly, all of the suppressor mutations in the RT/En domain map on one face of the RT/En domain. Five mutations of region e map within the RNaseH-like domain (Fig. 4a), and four of these are located within the β-finger, which forms a continuous β-sheet with the tail of Aar2 and the β-barrel in the Jab1/MPN domain. It is possible that this β-hairpin is also involved in a protein–protein interaction crucial for positioning the RNaseH-like domain in the U5 snRNP or spliceosome.

One of the Prp8-cat mutations (Val1098Asp) suppresses a cold-sensitive mutation in the first RecA domain in Brr2 known as *brr2-1* (Glu610Gly). The residues (1022–1214) covering the entire fingers/palm domain were subjected to further mutagenesis, and five other *brr2-1* suppressors were isolated[38] (Supplementary Table 7). These mutations also map within or at the base of the four-stranded β-sheet exposed on the surface of the reverse transcriptase palm domain (Fig. 4c). The exact mechanism of U4-cs1 suppression is unclear, but the fact that all suppressor mutations map on one face of the RT/En domain suggests that this is a crucial RNA-binding or, more likely, protein-binding surface. As *brr2-1* and some U4-cs1 suppressors map in the same region it may be a binding interface for Brr2 in addition to the Jab1/MPN domain. Brr2 bound on this surface would be ideally positioned to feed U6 snRNA, unwound from the U4/U6 duplex, into the active site cleft in Prp8 and may be in close proximity to the active site RNA. Ski2-type RNA helicases such as Brr2 unwind duplex RNA after loading onto a 3′ overhanging end. The extended helix I of the U4/U6 duplex in the U4-cs1 mutant alters the position and orientation of the 3′ overhanging nucleotides. It is possible that U4-cs1 and *brr2-1* suppressor mutations may slightly reposition Brr2 to facilitate loading of the U6/U4-cs1 substrate.

The origin of the spliceosome

The crystal structure of Prp8$^{885–2413}$ has revealed a new large domain consisting of reverse transcriptase and endonuclease domains. The fact that the palm and Th/X domains have considerable sequence and structural similarity to their counterparts in bacterial and fungal group II IEPs has very important functional and evolutionary implications. The farsighted hypothesis that nuclear pre-mRNA splicing and group II self-splicing have a common origin was based on the fact that in both processes introns are excised by two successive trans-esterification reactions by means of a lariat intermediate[39,40]. Self-splicing group II introns have six structurally conserved domains (domains I–VI) (ref. 41). The fact that the RNA core of the spliceosome contains structural and functional counterparts of domains V and VI and the exon-binding loop of group II intron further strengthened this hypothesis[9,39,40,42–44]. However, there is still an enormous

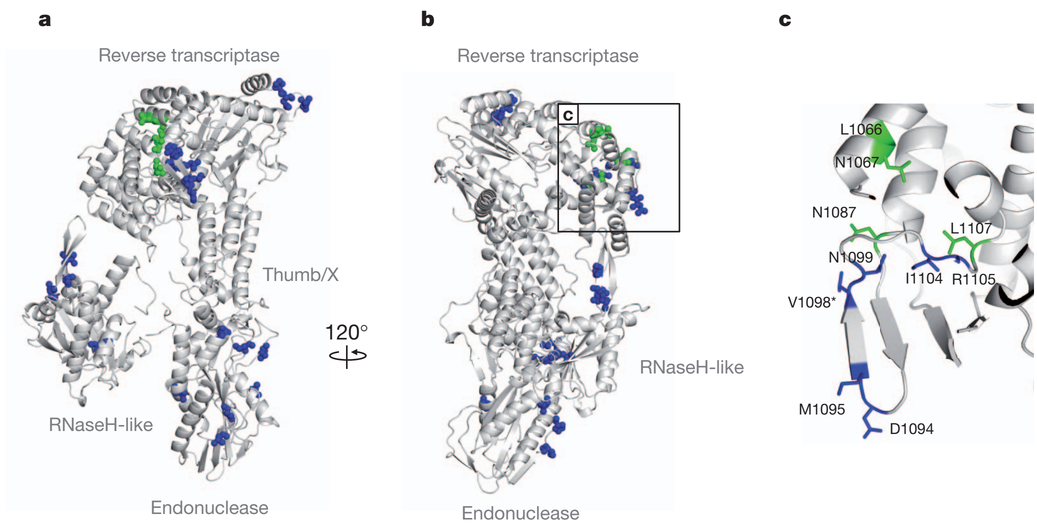


Figure 4 | Suppressors of U4-cs1 and *brr2-1* alleles mapped on the Prp8 structure. **a**, U4-cs-1 (blue spheres) and *brr2-1* (green spheres) suppressor mutants map on one face of the RT/En domain of Prp8. **b**, A view rotated by 120° along the *y* axis. **c**, Both types of suppressor mutant map to the same region of the Prp8 reverse transcriptase domain. Residues that suppress both alleles are marked with an asterisk.

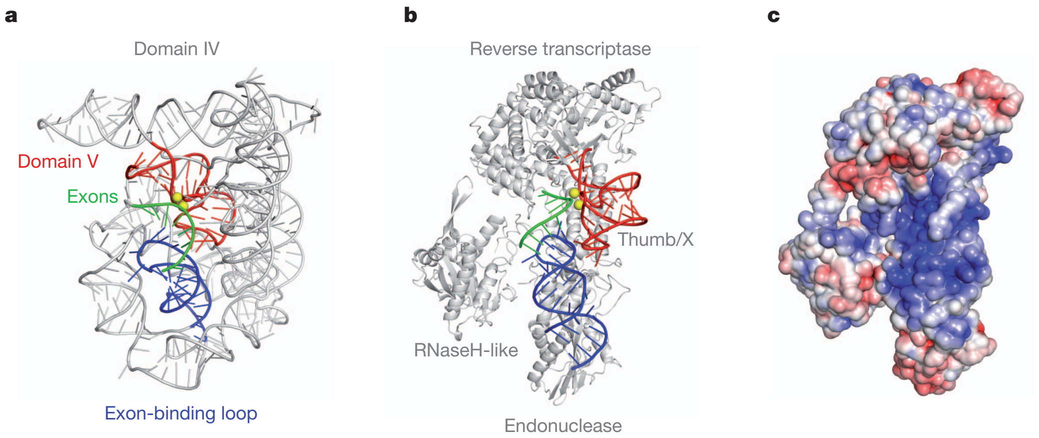


Figure 5 | Comparison between the active site of group II intron and the spliceosome (Prp8). **a**, Group II intron from *O. iheyensis* (PDB accession 3IGI). Domain V, red; exon-binding loop helix, blue; spliced exons, green; catalytic Mg^{2+} ions, yellow sphere; scaffolding RNA, grey. **b**, The RT/En domain with the RNaseH-like domain of Prp8 with the active RNA elements of group II intron modelled on its surface for size comparison. At present there are insufficient experimental constraints for the precise position or orientation of the RNA. **c**, Electrostatic potential ($\pm 5\,kTe^{-1}$) plotted on the solvent-accessible surface of the Prp8 (calculated with adaptive Poisson–Boltzmann solver, see Methods).

evolutionary gap between the two. Our crystal structure has, to our knowledge, provided the first experimental evidence for a link between a group II IEP and a component of the spliceosome. Group II introns are mobile genetic elements in which self-splicing is facilitated by the maturase activity of their IEP[43]. IEP remains bound to the excised intron and targets it to a homing site in genomic DNA where intron insertion is achieved by reverse splicing. The opposite DNA strand is then cleaved by the endonuclease domain of IEP and used as a primer for reverse transcription of the intron RNA by the IEP reverse transcriptase domain. IEP usually contains an endonuclease domain of the H-N-H endonuclease family[43], whereas Prp8 contains a type II restriction endonuclease domain.

Remarkably, organellar group II introns, split into independently transcribed segments[45], can undergo inefficient *trans*-splicing[41,45]. Ancestral nuclear pre-mRNA splicing activity could have evolved from the IEP open-reading frame (encoding ancestral Prp8) and the group II intron RNA domains (ancestral snRNAs), which became independent transcription units. The resulting incomplete group II introns could have been excised with the help of IEP and the *trans*-acting group II intron RNA domains and were gradually freed from the evolutionary constraints to maintain self-splicing activity as all of the RNA domains, except the branch-point sequence, became *trans*-acting elements. When some group II introns ceased to be mobile elements, the selective pressure to maintain the catalytically active reverse transcriptase domain was lost, but the reverse transcriptase domain continued to function as a maturase and became an assembly platform for the primitive snRNAs and substrate pre-mRNA. The reverse transcriptase domain, particularly the fingers and thumb domains, continued to evolve, and the addition of the RNaseH-like and Jab1/MPN domains facilitated the evolution of ancestral snRNAs into snRNPs as they recruited more proteins.

The crystal structure of an *Oceanobacillus iheyensis* group II intron reveals a tightly packed functional centre consisting of exon-binding loops, exons and domain V organized by the surrounding RNA scaffold[46,47]. No structure of a group II intron in complex with IEP has yet been reported, but the interaction between the *Lactococcus lactis* L1LtrB intron and its IEP has been studied biochemically[48–50]. The amino-terminal region of the reverse transcriptase domain binds intron DIVa with high affinity, whereas the Th/X domain makes contact with the catalytic core of the intron including the E1-DI, DII and DVI-E2 regions to promote splicing[48,50]. The functional RNA core of the spliceosome is postulated to be similar to that of a group II intron. Prp8 and the *O. iheyensis* group II intron have remarkably similar dimensions, and the Prp8 active site cavity is approximately the right size to accommodate the essential RNA domains of group II intron RNA (Fig. 5a, b). It is tempting to suggest that Prp8 has replaced the RNA scaffold surrounding the group II intron. Notably, the spliceosomal RNA catalytic core crosslinks to the region of Prp8 between reverse transcriptase and endonuclease domains (C. M. Norman and A.J.N., unpublished observations) and to the RNaseH-like domain[18] (Fig. 3b, c). The surface of this region exhibits extraordinary sequence conservation (Supplementary Fig. 18) and is remarkably electropositive (Fig. 5c). This is consistent with its role as the binding site for the RNA catalytic core. Structural analysis of a group II intron at different stages of catalysis has revealed that the intron active site can adopt two alternative conformations[46,47]. It has been proposed that Prp8 may undergo a transition between two alternative states that facilitate the first and second steps of splicing, respectively[35]. This transition may be achieved by repositioning of the RNaseH-like domain and the extended polypeptide chains (Fig. 3b, c), which line the inner surface of the active site cavity.

Until now, it has been hard to imagine how a ribonucleoprotein machine as immense and complex as the spliceosome could have evolved. The structure of Prp8 has given crucial insight into the active centre of the spliceosome, and its similarity to group II IEP provides a compelling link between group II self-splicing and the spliceosome.

METHODS SUMMARY

A large fragment of *S. cerevisiae* $Prp8^{885-2413}$ was co-expressed with Aar2 in yeast. The complex was purified by affinity chromatography using calmodulin-sepharose and Ni-NTA agarose, followed by ion-exchange chromatography on a mono-Q column. The complex was crystallized by the sitting-drop method. The structure was solved by molecular replacement using Aar2 and the RNaseH-like and Jab1/MPN domains of Prp8 as search models (PDB accessions 3SBT and 2OG4).

Full Methods and any associated references are available in the online version of the paper.

Received 2 November; accepted 18 December 2012.
Published online 23 January 2013.

1. Wahl, M. C., Will, C. L. & Lührmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701–718 (2009).
2. Wassarman, D. A. & Steitz, J. A. Interactions of small nuclear RNA's with precursor messenger RNA during *in vitro* splicing. *Science* **257,** 1918–1925 (1992).
3. Madhani, H. D. & Guthrie, C. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71,** 803–817 (1992).
4. Kandels-Lewis, S. & Seraphin, B. Involvement of U6 snRNA in 5′ splice site selection. *Science* **262,** 2035–2039 (1993).
5. Lesser, C. F. & Guthrie, C. Mutations in U6 snRNA that alter splice site specificity: Implications for the active site. *Science* **262,** 1982–1988 (1993).

6. Sun, J. S. & Manley, J. L. A novel U2–U6 snRNA structure is necessary for mammalian mRNA splicing. *Genes Dev.* **9,** 843–854 (1995).
7. Yean, S.-L., Wuenschell, G., Termini, J. & Lin, R. J. Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature* **408,** 881–884 (2000).
8. Steitz, T. A. & Steitz, J. A. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl Acad. Sci. USA* **90,** 6498–6502 (1993).
9. Newman, A. J. & Norman, C. U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. *Cell* **68,** 743–754 (1992).
10. Sontheimer, E. J. & Steitz, J. A. The U5 and U6 small nuclear RNAs as active site components of the spliceosome. *Science* **262,** 1989–1996 (1993).
11. O'Keefe, R. T., Norman, C. & Newman, A. J. The invariant U5 snRNA loop 1 sequence is dispensable for the first catalytic step of pre-mRNA splicing in yeast. *Cell* **86,** 679–689 (1996).
12. Achsel, T., Ahrens, K., Brahms, H., Teigelkamp, S. & Lührmann, R. The human U5-220kD protein (hPrp8) forms a stable RNA-free complex with several U5-specific proteins, including an unwindase, a homologue of ribosomal elongation factor EF-2, and a novel WD-40 protein. *Mol. Cell. Biol.* **18,** 6756–6766 (1998).
13. Bartels, C., Urlaub, H., Lührmann, R. & Fabrizio, P. Mutagenesis suggests several roles of Snu114p in pre-mRNA splicing. *J. Biol. Chem.* **278,** 28324–28334 (2003).
14. Small, E. C., Leggett, S. R., Winans, A. A. & Staley, J. P. The EF-G-like GTPase Snu114p regulates spliceosome dynamics mediated by Brr2p, a DExD/H box ATPase. *Mol. Cell* **23,** 389–399 (2006).
15. Teigelkamp, S., Newman, A. J. & Beggs, J. D. Extensive interactions of PRP8 protein with the 5′ and 3′ splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA. *EMBO J.* **14,** 2602–2612 (1995).
16. Dix, I., Russell, C. S., O'Keefe, R. T., Newman, A. J. & Beggs, J. D. Protein-RNA interactions in the U5 snRNP of *Saccharomyces cerevisiae. RNA* **4,** 1239–1250 (1998).
17. Vidal, V. P., Verdone, L., Mayes, A. E. & Beggs, J. D. Characterization of U6 snRNA-protein interactions. *RNA* **5,** 1470–1481 (1999).
18. Reyes, J. L., Gustafson, E. H., Luo, H. R., Moore, M. J. & Konarska, M. M. The C-terminal region of hPrp8 interacts with the conserved GU dinucleotide at the 5′ splice site. *RNA* **5,** 167–179 (1999).
19. MacMillan, A. M. *et al.* Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev.* **8,** 3008–3020 (1994).
20. Turner, I. A., Norman, C. M., Churcher, M. J. & Newman, A. J. Dissection of Prp8 protein defines multiple interactions with crucial RNA sequences in the catalytic core of the spliceosome. *RNA* **12,** 375–386 (2006).
21. Grainger, R. J. & Beggs, J. D. Prp8 protein: at the heart of the spliceosome. *RNA* **11,** 533–557 (2005).
22. Pena, V., Rozov, A., Fabrizio, P., Lührmann, R. & Wahl, M. C. Structure and function of an RNase H domain at the heart of the spliceosome. *EMBO J.* **27,** 2929–2940 (2008).
23. Ritchie, D. B. *et al.* Structural elucidation of a PRP8 core domain from the heart of the spliceosome. *Nature Struct. Mol. Biol.* **15,** 1199–1205 (2008).
24. Yang, K., Zhang, L., Xu, T., Heroux, A. & Zhao, R. Crystal structure of the β-finger domain of Prp8 reveals analogy to ribosomal proteins. *Proc. Natl Acad. Sci. USA* **105,** 13817–13822 (2008).
25. Pena, V., Liu, S., Bujnicki, J. M., Lührmann, R. & Wahl, M. C. Structure of a multipartite protein-protein interaction domain in splicing factor Prp8 and its link to *Retinitis pigmentosa*. *Mol. Cell* **25,** 615–624 (2007).
26. Zhang, L. *et al.* Crystal structure of the C-terminal domain of splicing factor Prp8 carrying retinitis pigmentosa mutants. *Protein Sci.* **16,** 1024–1031 (2007).
27. Dlakić, M. & Mushegian, A. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA* **17,** 799–808 (2011).
28. Boon, K. L. *et al.* Prp8 mutations that cause human retinitis pigmentosa lead to a U5 snRNP maturation defect in yeast. *Nature Struct. Mol. Biol.* **14,** 1077–1083 (2007).
29. Weber, G. *et al.* Mechanism for Aar2p function as a U5 snRNP assembly factor. *Genes Dev.* **25,** 1601–1612 (2011).
30. Joyce, C. M. & Steitz, T. A. Function and structure relationships in DNA polymerase. *Annu. Rev. Biochem.* **63,** 777–822 (1994).
31. Dias, A. *et al.* The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458,** 914–918 (2009).
32. Yuan, P. *et al.* Crystal structure of an avian influenza polymerase PA_N reveals an endonuclease active site. *Nature* **458,** 909–913 (2009).
33. Wyatt, J. R., Sontheimer, E. J. & Steitz, J. A. Site-specific cross-linking of mammalian U5 snRNP to the 5′ splice site before the first step of pre-mRNA splicing. *Genes Dev.* **6,** 2542–2553 (1992).
34. Urlaub, H., Hartmuth, K., Kostka, S., Grelle, G. & Lührmann, R. A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). *J. Biol. Chem.* **275,** 41458–41468 (2000).
35. Query, C. C. & Konarska, M. M. Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. *Mol. Cell* **14,** 343–354 (2004).
36. Umen, J. G. & Guthrie, C. Mutagenesis of the yeast gene PRP8 reveals domains governing the specificity and fidelity of 3′ splice site selection. *Genetics* **143,** 723–739 (1996).
37. Kuhn, A. N. & Brow, D. A. Suppressors of a cold-sensitive mutation in yeast U4 RNA define five domains in the splicing factor Prp8 that influence spliceosome activation. *Genetics* **155,** 1667–1682 (2000).
38. Kuhn, A. N., Reichl, E. M. & Brow, D. A. Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation. *Proc. Natl Acad. Sci. USA* **99,** 9145–9149 (2002).
39. Sharp, P. A. On the origin of RNA splicing and introns. *Cell* **42,** 397–400 (1985).
40. Cech, T. R. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* **44,** 207–210 (1986).
41. Michel, F., Umesono, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82,** 5–30 (1989).
42. Sharp, P. A. Five easy pieces. *Science* **254,** 663 (1991).
43. Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annu. Rev. Genet.* **38,** 1–35 (2004).
44. Pyle, A. M. & Lambowitz, A. M. in *The RNA World* 3rd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 469–505 (Cold Spring Harbor Laboratory Press, 2006).
45. Qui, Y.-L. & Palmer, J. D. Many different origins of trans splicing in a plant mitochondrial group II intron. *J. Mol. Evol.* **59,** 80–89 (2004).
46. Toor, N. *et al.* Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. *RNA* **16,** 57–69 (2010).
47. Marcia, M. & Pyle, A. M. Visualizing group II intron catalysis through the stages of splicing. *Cell* **151,** 497–507 (2012).
48. Matsuura, M., Noah, J. W. & Lambowitz, A. M. Mechanism of maturase-promoted group II intron splicing. *EMBO J.* **20,** 7259–7270 (2001).
49. Rambo, R. P. & Doudna, J. A. Assembly of an active group II intron-maturase complex by protein dimerization. *Biochemistry* **43,** 6486–6497 (2004).
50. Gu, S. Q. *et al.* Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA* **16,** 732–747 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Murshudov for his help and guidance with crystallography; T.H.D. Nguyen, Y. Kondo, M. van Roon, J. Hardin, J. Li, C. Norman, A. Andreeva, A. Murzin and M. Yu for discussion and help; T. Ignjatovic and H. Oshikane for their contributions at the early stage of the project; L. Passmore and L. Jovine for reading of the manuscript; and M. Ikura for the gift of a calmodulin clone. We are grateful to the beamline staff at Diamond Light Source and European Synchrotron Radiation Facility for their help, and to E. Stephens and the LMB mass spectrometry facility for their help. W.P.G. thanks the Cambridge European Trust and Downing College for scholarships. This project was funded by the UK Medical Research Council.

Author Contributions A.J.N. and K.N. initiated the project and worked on protein expression and purification for many years. Co-expression of Prp8 and Aar2 by A.J.N. was a crucial step of the project. W.P.G. successfully identified and expressed a stable large fragment of Prp8, crystallized the Prp8–Aar2 complex and solved and refined the structure almost single-handedly with practical support from K.N. and A.J.N. C.O. analysed the mercury derivative data and refined the structure of the $P2_12_12_1$ crystal form. W.P.G. and K.N. analysed the structure and wrote the paper with important input from A.J.N. and C.O.

Author Information Atomic coordinates and structure factors for the Prp8–Aar2 complex have been deposited in the Protein Data Bank under accession codes 4I43 ($C222_1$) and 3ZEF ($P2_12_12_1$). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.N. (kn@mrc-lmb.cam.ac.uk) or A.J.N. (newman@mrc-lmb.cam.ac.uk).

METHODS

Protein expression and purification. The Prp8$^{885-2413}$ fragment fused to an N-terminal calmodulin-binding peptide and a full-length Aar2 with a C-terminal 8×His-tag were cloned into pUC18 vectors containing the expression cassette described previously[51]. These expression cassettes were transferred to pRS426 (Prp8) and pRS424 (Aar2) plasmids[52]. BCY123 cells (*MATα pep4::HIS3 prb1::LEU2 bar1::HIS6 lys2::GAL1/10-GAL4 can1 ade2 trp1 ura3 his3 leu2-3,112*) containing both plasmids were grown on selective medium (lacking uracil and tryptophan) with 1% raffinose to $A_{600\,nm} = 0.8$–1.0. Protein expression was induced by the addition of galactose to the final concentration of 2% and cells were grown at 30 °C for 12–16 h. Cell pellets were resuspended in one volume of 2× CAL350 buffer (700 mM NaCl, 100 mM Tris-HCl, pH 8.5, 4 mM $CaCl_2$, 2 mM Mg-acetate, 2 mM imidazole, 20 mM 2-mercaptoethanol, 0.2% Igepal CA-630, EDTA-free protease inhibitor cocktail (Roche)), and frozen in liquid nitrogen. Solid-phase cell disruption was performed with freezer mill 6870 (SPEX CertiPrep) and the crude extract was adjusted to pH 8.0 with Tris base, then centrifuged at 48,000*g* at 4 °C for 30 min. The supernatant was incubated with calmodulin-sepharose (recombinant calmodulin coupled to cyanogen bromide-activated sepharose (GE)) overnight at 4 °C. Resin was washed with CAL500W buffer (500 mM NaCl, 20 mM Tris-HCl, pH 8.0, 2 mM $CaCl_2$, 1 mM Mg-acetate, 1 mM imidazole, 10 mM 2-mercaptoethanol) and eluted with CAL500E buffer (500 mM NaCl, 20 mM Tris-HCl, pH 8.0, 2 mM EGTA, 1 mM Mg-acetate, 1 mM imidazole, 10 mM 2-mercaptoethanol). After dialysis against Ni-dialysis buffer (500 mM NaCl, 20 mM Tris-HCl, pH 8.0, 10 mM 2-mercaptoethanol, 5 mM imidazole) at 4 °C, the sample was incubated with Ni-NTA agarose for 3–6 h. The Ni-NTA agarose was packed into a small column, washed with Ni500W buffer (500 mM NaCl, 20 mM Tris-HCl, pH 8.0, 20 mM imidazole, 10 mM 2-mercaptoethanol) and the protein was eluted with Ni500E buffer (500 mM NaCl, 20 mM Tris-HCl, pH 8.0, 250 mM imidazole, 10 mM 2-mercaptoethanol). The eluate was dialysed against: 300 mM KCl, 20 mM K-HEPES, pH 7.8, 1 mM dithiothreitol (DTT), and diluted with one-third volume of: 20 mM K-HEPES, pH 7.8, 1 mM DTT. The sample was applied to a MonoQ ion exchange column (10/100 GL) equilibrated with a 0.85:0.15 mixture of buffer A (50 mM KCl, 20 mM K-HEPES, pH 7.8, 1 mM DTT) and buffer B (1 M KCl, 20 mM K-HEPES, pH 7.8, 1 mM DTT) and eluted with a linear gradient of 15–50% of buffer B over ten column volumes. Typically, a 24-l culture yielded 4–8 mg of purified protein.

Crystallization. Crystals of the Prp8$^{885-2413}$–Aar2 complex were obtained by sitting-drop vapour diffusion technique at 293 K. The protein solution (10–25 mg ml^{-1} in 300 mM KCl, 20 mM K-HEPES, pH 7.8, 1 mM DTT) was mixed with an equal amount of reservoir solution (7–9% PEG 8000, 100 mM sodium citrate, 50–200 mM ammonium sulphate) and equilibrated against reservoir solution for 1 h before streak-seeding with a feline whisker. Crystals suitable for data collection appeared within 1–3 days, reaching maximum dimensions of $0.6 \times 0.2 \times 0.1\ mm^3$. Cryo-protection was achieved in three steps by in-well buffer exchange with (1) 20% PEG 8000, 100 mM sodium citrate, for 5 min; (2) 30% PEG 8000, 100 mM sodium citrate, for 5 min; and (3) 30% PEG 8000, 5% PEG 400, 100 mM sodium citrate, for 60 min. For mercury derivative crystals, the last step of cryo-protection was extended to 16 h and the buffer was supplemented with 1 mM methylmercury nitrate. Cryo-protected crystals were flash frozen in liquid nitrogen for data collection.

Crystals in both space groups (Supplementary Table 1) appeared under the same condition. Initial diffraction of the mercury derivative crystals was limited to 8–10 Å, regardless of the concentration of the mercury compound, soaking time and which mercury compound was used. Close inspection of the 3SBT crystal structure[29] revealed two cysteines in Aar2 (C251 and C292) that were in close proximity (~3.5 Å apart). Successful derivatization of both cysteines would certainly lead to a steric clash and subsequent disintegration of the crystal lattice, which could explain the observed loss of diffraction. An Aar2 double mutant (C251S/C292S) was produced to overcome this limitation. Mutant protein crystallized under the same condition and the diffraction of the derivative crystals was improved from 8 to 3.7 Å.

Data collection and processing. Data for the $C222_1$ crystals was collected using an ADSC Q315R detector at ESRF (beamline ID23-1) at 0.91 Å wavelength with 0.1° oscillation range. Data for the $P2_12_12_1$ crystals were collected at the Diamond Light Source (beamlines I02 and I03) with a Pilatus 6 M detector. A mercury derivative single-wavelength anomalous dispersion (SAD) data set was collected at the Hg LIII peak wavelength (1.00726 Å) with 0.2° oscillation range. Images were indexed and integrated in Mosflm[53] or XDS[54] and then scaled and analysed in Aimless[55].

Structure determination. An initial solution for the $C222_1$ crystals was obtained by molecular replacement with Phaser[56] and Molrep[57]. Search models were obtained from the PDB: Jab1/MPN domain, accession 2OG4; RNaseH-like domain and Aar2, accession 3SBT. Domains were located one-by-one in the following order: Aar2, RNaseH and Jab1/MPN. The solution was verified by a comparison with anomalous density peak positions in methylmercury nitrate derivative data in the $P2_12_12_1$ space group (Supplementary Fig. 1). Molecular replacement solution was refined in Refmac5 (ref. 58) (100 cycles of jelly-body refinement) and boosted by 50 cycles of automated model building/density modification in SHELXE[59]. The resulting improved phases were used to extend the existing model in ARP/wARP 7.3 (ref. 60). This procedure was repeated twice. Structure was manually rebuilt and modified in Coot[61] and refined in Refmac5. Structure validation was performed with Coot and Molprobity server[62].

The $P2_12_12_1$ crystal form was solved by molecular replacement with Phaser using the refined $C222_1$ structure as a search model. The density map from Phaser indicated that some regions of the Prp8$^{885-2413}$–Aar2 complex that were ordered in the $C222_1$ crystals were disordered in the $P2_12_12_1$ crystals and that other parts of the complex were orientated a little differently, particularly the Jab/MPN and RNaseH-like domains. The structure was modified by real-space rigid body refinement and manual rebuilding in Coot and refined in Refmac5 (ref. 58).

Structure analysis. Structural similarity between the Prp8$^{885-2413}$–Aar2 complex and other proteins in PDB databases was assessed by secondary structure matching as implemented in the PDBeFold web server[63]. Pairwise superpositions were made using the DaliLite server[64]. Surface electrostatic potential was calculated with adaptive Poisson–Boltzmann solver[65] implemented in Pymol. Surface conservation analysis was performed with ConSurf web server[66]. Structure was visualized in Pymol (http://www.pymol.org). Multiple sequence alignment was prepared with ClustalW[67]. Secondary structure assignment was carried out using DSSP[68] and STRIDE[69].

Plasmid shuffling. Viability of the Prp8 mutants was assessed by the plasmid shuffling method. Prp8 deletion-mutant strain (SC261Δ8B1)[19], carrying wild-type PRP8 on pRS316 (URA3, centromeric replication origin)[70] was transformed with mutant Prp8 on pRS314 (TRP1, centromeric replication origin)[70] and transformants were selected on plates lacking tryptophan. Cells were transferred onto plates containing 5-fluoro-orotic acid (5-FOA), to test cell growth after loss of the uracil plasmid. Viability was assessed visually after 3–5 days of incubation at 30 °C. Plasmids from the 5-FOA-resistant strains were rescued and sequenced to eliminate possibility of the observed phenotype being a result of a recombination event at the earlier stages of experiment.

51. Wagenbach, M. *et al.* Synthesis of wild type and mutant human hemoglobins in *Saccharomyces cerevisiae*. *Biotechnology (N Y)* **9,** 57–61 (1991).
52. Christianson, T. W., Sikorski, R. S., Dante, M., Shero, J. H. & Hieter, P. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110,** 119–122 (1992).
53. Leslie, A. G. W. & Powell, H. R. Processing diffraction data with Mosflm. *Evolv. Methods Macromol. Crystallograph.* **245,** 41–51 (2007).
54. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
55. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
56. McCoy, A. J. *et al. Phaser* crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
57. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66,** 22–25 (2010).
58. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
59. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66,** 479–485 (2010).
60. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3,** 1171–1179 (2008).
61. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
62. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
63. Krissinel, E. & Henrick, K. Secondary-structure matching (PDBeFold), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* **60,** 2256–2268 (2004).
64. Holm, L. & Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics* **16,** 566–567 (2000).
65. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).
66. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** 529–533 (2010).
67. Larkin, M. A. *et al.* ClustalW and ClustalX version 2. *Bioinformatics* **23,** 2947–2948 (2007).
68. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).
69. Heinig, M. & Frishman, D. STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32,** W500–W502 (2004).
70. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122,** 19–27 (1989).

doi:10.1038/nature11805

An old disk still capable of forming a planetary system

Edwin A. Bergin[1], L. Ilsedore Cleeves[1], Uma Gorti[2,3], Ke Zhang[4], Geoffrey A. Blake[5], Joel D. Green[6], Sean M. Andrews[7], Neal J. Evans II[6], Thomas Henning[8], Karin Öberg[7], Klaus Pontoppidan[9], Chunhua Qi[7], Colette Salyk[10] & Ewine F. van Dishoeck[11,12]

From the masses of the planets orbiting the Sun, and the abundance of elements relative to hydrogen, it is estimated that when the Solar System formed, the circumstellar disk must have had a minimum mass of around 0.01 solar masses within about 100 astronomical units of the star[1–4]. (One astronomical unit is the Earth–Sun distance.) The main constituent of the disk, gaseous molecular hydrogen, does not efficiently emit radiation from the disk mass reservoir[5], and so the most common measure of the disk mass is dust thermal emission and lines of gaseous carbon monoxide[6]. Carbon monoxide emission generally indicates properties of the disk surface, and the conversion from dust emission to gas mass requires knowledge of the grain properties and the gas-to-dust mass ratio, which probably differ from their interstellar values[7,8]. As a result, mass estimates vary by orders of magnitude, as exemplified by the relatively old (3–10 million years) star TW Hydrae[9,10], for which the range is 0.0005–0.06 solar masses[11–14]. Here we report the detection of the fundamental rotational transition of hydrogen deuteride from the direction of TW Hydrae. Hydrogen deuteride is a good tracer of disk gas because it follows the distribution of molecular hydrogen and its emission is sensitive to the total mass. The detection of hydrogen deuteride, combined with existing observations and detailed models, implies a disk mass of more than 0.05 solar masses, which is enough to form a planetary system like our own.

Commonly used tracers of protoplanetary disk masses are thermal emission from dust grains and rotational lines of carbon monoxide (CO) gas. However, the methods by which these are detected rely on unconstrained assumptions. The dust detection method has to assume an opacity per gram of dust, and grain growth can change this value drastically[15]. The gas mass is then calculated by multiplying the dust mass by the gas-to-dust ratio, which is usually assumed to be $\sim$100 from measurements of the interstellar medium[16]. The gas mass thus depends on a large and uncertain correction factor. The alternative is to use rotational CO lines as gas tracers, but their emission is optically thick and therefore trace the disk surface temperature rather than the midplane mass. The use of CO as a gas tracer thus leads to large discrepancies between mass estimates for different models of TW Hya (from $5 \times 10^{-4} M_\odot$ to $0.06 M_\odot$, where $M_\odot$ is the solar mass), even though each matches a similar set of observations[13,14].

Using the Herschel Space Observatory[17] Photodetector Array Camera and Spectrometer[18], we robustly detected (9σ) the lowest rotational transition, $J = 1 \rightarrow 0$, of hydrogen deuteride (HD) in the closest ($D \approx 55$ pc) and best-studied circumstellar disk around TW Hya (Fig. 1). This star is older (3–10 Myr; refs 9, 10, 19) than most stars with gas-rich circumstellar disks[8]. The abundance of deuterium atoms relative to hydrogen is well characterized, via atomic electronic transitions, to be $x_\mathrm{D} = (1.5 \pm 0.1) \times 10^{-5}$ in objects that reside within $\sim$100 pc of the Sun[20]. Adding a hydrogen atom to each, to form H_2 and HD, which is appropriate for much of the disk mass, provides an HD abundance relative to H_2 of $x_\mathrm{HD} = 3.0 \times 10^{-5}$. We combine the HD data with existing molecular observations to set new constraints on the disk mass within 100 AU, which is the most fundamental quantity that determines whether planets can form. The disk mass also

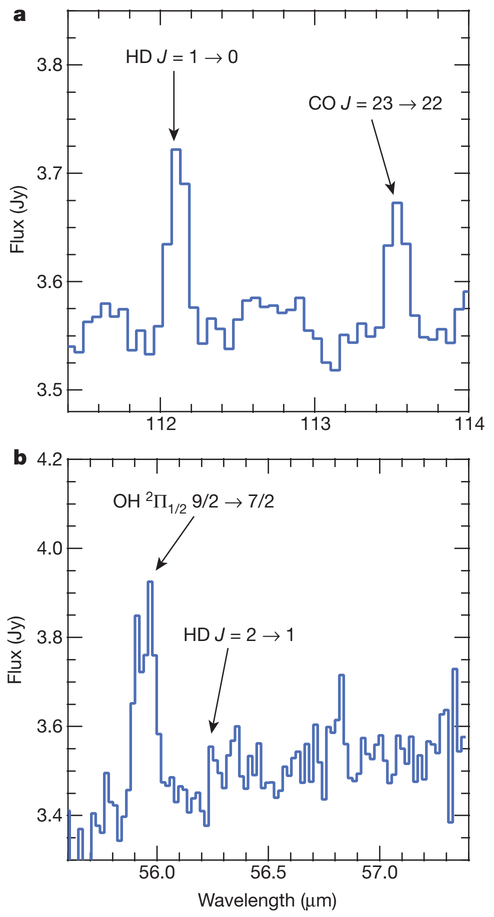


Figure 1 | Herschel detection of HD in the TW Hya protoplanetary disk. **a**, The fundamental $J = 1 \rightarrow 0$ line of HD lies at $\sim$112 μm. On 20 November 2011, it was detected from the direction of the TW Hya disk at the 9σ level. The total integrated flux is $(6.3 \pm 0.7) \times 10^{-18}$ W m^{-2}. We also report a detection of the warm disk atmosphere in CO $J = 23 \rightarrow 22$ with a total integrated flux of $(4.4 \pm 0.7) \times 10^{-18}$ W m^{-2}. The $J = 1 \rightarrow 0$ line of HD was previously detected by the Infrared Space Observatory in a warm gas cloud exposed to radiation from nearby stars[27]. Other transitions have also been detected in shocked regions associated with supernovae and outflows from massive stars[28,29]. **b**, Simultaneous observations of HD $J = 2 \rightarrow 1$ are shown. For HD $J = 2 \rightarrow 1$, we find a detection limit of $< 8.0 \times 10^{-18}$ W m^{-2} (3σ). We also report a detection of the OH $^2\Pi_{1/2}$ 9/2 → 7/2 doublet near 55.94 μm with an integrated flux of $(4.93 \pm 0.27) \times 10^{-17}$ W m^{-2}. The spectra include the observed thermal dust continuum of $\sim$ 3.55 Jy at both wavelengths.

[1]Department of Astronomy, University of Michigan, 500 Church Street, Ann Arbor, Michigan 48109, USA. [2]SETI Institute, Mountain View, California 94043, USA. [3]NASA Ames Research Center, Moffett Field, California 94035, USA. [4]California Institute of Technology, Division of Physics, Mathematics and Astronomy, MS 150-21, Pasadena, California 91125, USA. [5]California Institute of Technology, Division of Geological and Planetary Sciences, MS 150-21, Pasadena, California 91125, USA. [6]Department of Astronomy, The University of Texas, 2515 Speedway, Stop C1402, Austin, Texas 78712, USA. [7]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [8]Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany. [9]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [10]National Optical Astronomy Observatory, 950 North Cherry Avenue, Tucson, Arizona 85719, USA. [11]Max Planck Institut für Extraterrestrische Physik, Giessenbachstrasse 1, 85748 Garching, Germany. [12]Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands.

governs the primary mode of giant-planet formation, either through core accretion or gravitational instability[21]. In this context, we do not know whether the Solar System formed within a typical disk, because nearly half of the present estimates of extrasolar disk masses are less than the minimum solar nebula mass[8]. Our current census of extrasolar planetary systems furthermore suggests that even larger disk masses are necessary to form many of the exoplanetary systems seen[22,23].

With smaller rotational energy spacings and a weak electric dipole moment, HD $J=1\rightarrow 0$ is one million times more emissive than H_2 for a given gas mass at a gas temperature of $T_{gas}=20$ K. The HD line flux (F_l) sets a lower limit to the H_2 gas mass at distance D (Supplementary Information):

$$M_{\rm gasdisk} > 5.2\times10^{-5}\left(\frac{F_{\rm l}}{6.3\times10^{-18}\ {\rm W\ m^{-2}}}\right)\left(\frac{3\times10^{-5}}{x_{\rm HD}}\right) \times\left(\frac{D}{55\ {\rm pc}}\right)^2\exp\left(\frac{128.5\ {\rm K}}{T_{\rm gas}}\right)M_\odot \quad (1)$$

If HD is optically thick or deuterium is contained in other molecules such as polycyclic aromatic hydrocarbons or molecular ices, the conversion from deuterium mass to hydrogen mass will be higher and the mass will thus be larger, hence the lower limit. The strong temperature dependence arises from the fractional population of the $J=1$ state, which has a value of $f_{J=1}\approx 3\exp(-128.5\,{\rm K}/T_{\rm gas})$ for $T_{\rm gas}<50$ K in thermal equilibrium. Owing to the low fractional population in the $J=1$ state, HD does not emit appreciably from gas with $T\approx 10$–15 K, which is the estimated temperature in the outer disk mass reservoir (at a radius $R\gtrsim 20$–40 AU). The HD mass derived from equation (1) provides an estimate of the mass in warm gas, and is therefore a lower limit on the total mass within 100 AU.

The only factor in equation (1) that could lower the mass estimate is a higher $T_{\rm gas}$. The upper limit on the $J=2\rightarrow 1$ transition of HD (Fig. 1) implies that $T_{\rm gas}<80$ K in the emitting region. This $T_{\rm gas}$ estimate yields $M_{\rm gas\ disk}>2.2\times10^{-4}M_\odot$, but $T_{\rm gas}$ is unlikely to be this high for the bulk of the disk. CO rotational transitions are optically thick and the level populations are in equilibrium with $T_{\rm gas}$, and so they provide a measure of $T_{\rm gas}$. Atacama Large Millimeter/submillimeter Array (ALMA) observations of CO $J=3\rightarrow 2$ emission in a $1.7''\times1.5''$ beam (corresponding to gas within a radius of ~43 AU) (Supplementary Information and Supplementary Fig. 1) yield an average $T_{\rm gas}$ of 29.7 K within 43 AU, and $M_{\rm gas\ disk}>3.9\times10^{-3}M_\odot$. This value is still likely to be too low, because the emission from optically thick CO presumably gives information about material closer to the surface than does HD, and this gas will be warmer than the HD line-emitting region. Thus, essentially all correction factors would increase the mass beyond this conservative limit, which already rules out a portion of the low end of previous mass determinations.

To determine the mass more accurately, we turn to detailed models that incorporate explicit gas thermal physics providing for substantial radial and vertical thermal structure. Both published models of the TW Hya disk reproduce a range of gas-phase emission lines, but in one case with $M_{\rm gas\ disk}=0.06M_\odot$ (ref. 14) and in the other with $M_{\rm gas\ disk}=0.003M_\odot$ (ref. 13) (Supplementary Information and Supplementary Table 1). These models were both placed into detailed radiation transfer simulations. The results from this calculation and the adopted physical structure are given in Fig. 2 for the model with $M_{\rm gas\ disk}=0.06M_\odot$. Figure 2c shows the cumulative flux as a function of radius for the higher-mass model; over 80% of the emission is predicted to arise from gas within a radius of 80 AU. Furthermore, Fig. 2d provides a calculation of the HD emissive mass as a function of gas temperature. This calculation suggests that gas with a temperature of 30–50 K is responsible for the majority of the HD emission.

The model with $M_{\rm gas\ disk}=0.003M_\odot$ predicts an HD line flux of $F_{\rm l}=3.8\times10^{-19}$ W m^{-2}, which is more than an order of magnitude below the detected level. For this model to reach the observed flux, the

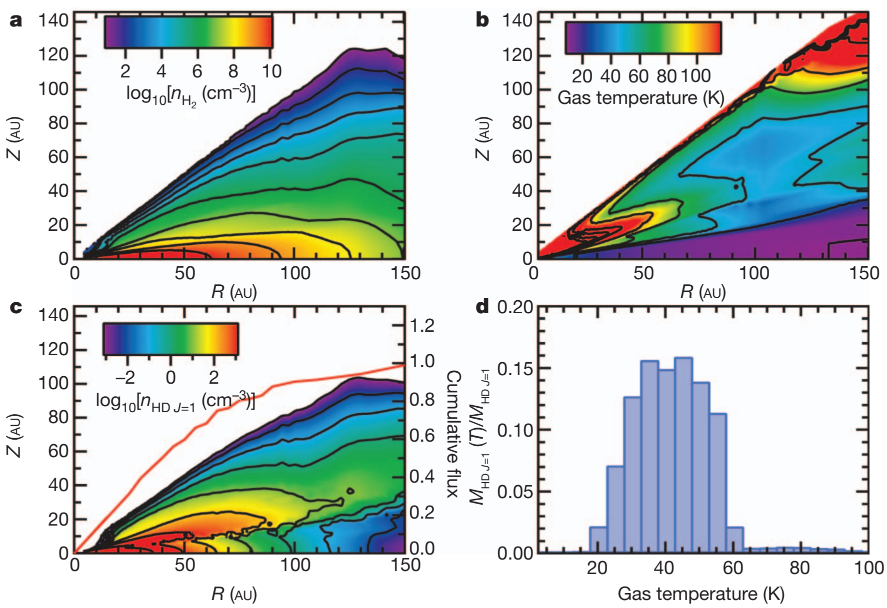


Figure 2 | Model of the physical structure and HD emission of the TW Hya circumstellar disk. **a**, Radial (R) and vertical (Z) distribution of the H_2 volume density, n_{H_2}, calculated in a model disk with mass $0.06M_\odot$ (ref. 14). Contours start from the top at $\log_{10}[n_{H_2}\ ({\rm cm^{-3}})]=1.0$ and are stepped in units of factors of ten. **b**, Gas temperature structure as derived by the thermochemical model[14]. Contours are at 10, 25, 50, 75, 100, 150, 200, 250 and 300 K. **c**, Radial and vertical distribution of the HD $J=1$ volume density, $n_{{\rm HD}\,J=1}$, predicted in a model disk with the gas density and temperature structure as given in **a** and **b**, with an HD abundance relative to H_2 of 3.0×10^{-5}. Contours start from the top at $\log_{10}[n_{{\rm HD}\,J=1}\ ({\rm cm^{-3}})]=-3$ and are stepped in factors of ten. The red line shows the cumulative flux contribution as a function of radius in terms of fractions of the overall predicted flux, 3.1×10^{-18} W m^{-2}. To predict the HD line emission, we calculate the solution of the equations of statistical equilibrium including the effects of line and dust opacity using the LIME code[30]. **d**, Fraction of the HD emission arising from gas with different temperatures, computed as a function of the mass of HD excited to the $J=1$ state in gas at temperatures binned in units of 5 K ($M_{{\rm HD}\,J=1}(T)$) normalized to the total mass of HD with $J=1$ ($M_{{\rm HD}\,J=1}$). In particular, $\int n_{{\rm HD}\,J=1}2\pi R\,dr\,dz$ is computed successively in gas temperature bins of 5 K and then normalized to the total mass of HD in the $J=1$ state.

disk mass would have to be 20 times greater and so this lower mass is ruled out. The $M_{gasdisk} = 0.06 M_{\odot}$ model predicts that $F_l = 3.1 \times 10^{-18}$ W m^{-2}, which is still a factor of two below the observed value: even the 'high' mass estimate is too low. On the basis of this model, we estimate that the disk gas mass within 80 AU, where the majority of HD emissions arise, is $0.056 M_{\odot}$. Both of these models match other observations: the low-disk-mass model matches CO and ^{13}CO $J = 3 \rightarrow 2$ emission, and the higher-mass model reproduces CO $J = 2 \rightarrow 1$, $J = 3 \rightarrow 2$ and $J = 6 \rightarrow 5$ emission. Both models also reproduce observed emission from other species. However, they differ by a factor of ten in predicting the HD emission. This difference shows the value of HD in constraining masses.

The age of TW Hya is uncertain. The canonical age of the cluster is 10^{+10}_{-7} Myr (ref. 9). However, there could be an age spread in cluster members, and ages estimated for TW Hya itself range from 3 to 10 Myr (refs 10, 19). Even at the low end of this range, TW Hya is older than the half-life of gaseous disks, which is inferred to be about 2 Myr (ref. 8). In the case of the TW Hya association, there is also little evidence for an associated molecular cloud[24], which is an additional indicator that this system is relatively older than most gas-rich disks. The lifetime of the gaseous disk is important because it sets the available time frame for the formation of gas giants equivalent to Jupiter or Saturn. According to our analysis, TW Hya contains a massive gas disk ($\gtrsim 0.06 M_{\odot}$) that is several times the minimum mass required to make the planets in the Solar System. Thus, this 'old' disk can still form a planetary system like our own.

The recent detection of cold water vapour from TW Hya yielded indirect evidence for a large water-ice reservoir (equal in mass to several thousand Earth oceans) assuming a disk mass of $0.02 M_{\odot}$ (ref. 25). Our higher mass estimate implies a larger water-ice reservoir, perhaps greater in mass by a factor of two. The mass estimate in this system lies at the upper end of previous mass measurements[8], hinting that other disk masses are underestimated. The main uncertainty in the masses derived here is the gas temperature structure of the disk. In future, observations of optically thick molecular lines, particularly CO, can be used to trace the thermal structure of gas in the disk. Observations of rarer CO isotopologues will then provide constraints on the temperature in deeper layers[26]. With ALMA, we will readily resolve multiple gas temperature tracers within a radius of 80 AU, where HD strongly emits. When these are used in tandem with HD, we will be able to derive the gas mass with much greater accuracy (our simulations suggest to within a factor of 2–3). Moreover, additional HD detections could be provided by the Herschel Space Observatory and with higher spectral resolution by the German Receiver for Astronomy at Terahertz Frequencies on board the Stratospheric Observatory for Infrared Astronomy under favourable atmospheric conditions. These data could be used alongside emission from species such as $C^{18}O$, $C^{17}O$ or the dust to calibrate these more widely available probes to determine the disk gas mass. Thus, with the use of HD to complement other observations and constrain models, we may finally place useful constraints on one of the most important quantities that governs the process of planetary formation.

Received 21 June; accepted 14 November 2012.

1. Kuiper, G. P. The formation of the planets, part III. *J. R. Astron. Soc. Can.* **50,** 158–176 (1956).
2. Kusaka, T., Nakano, T. & Hayashi, C. Growth of solid particles in the primordial solar nebula. *Prog. Theor. Phys.* **44,** 1580–1595 (1970).
3. Weidenschilling, S. J. Aerodynamics of solid bodies in the solar nebula. *Mon. Not. R. Astron. Soc.* **180,** 57–70 (1977).
4. Hayashi, C. Structure of the solar nebula, growth and decay of magnetic fields and effects of magnetic and turbulent viscosities on the nebula. *Prog. Theor. Phys.* **70** (suppl.), 35–53 (1981).
5. Carmona, A. *et al.* A search for mid-infrared molecular hydrogen emission from protoplanetary disks. *Astron. Astrophys.* **477,** 839–852 (2008).
6. Beckwith, S. V. W., Sargent, A. I., Chini, R. S. & Guesten, R. A survey for circumstellar disks around young stellar objects. *Astron. J.* **99,** 924–945 (1990).
7. Hartmann, L. Masses and mass distributions of protoplanetary disks. *Phys. Scripta* 014012 (2008).
8. Williams, J. P. & Cieza, L. A. Protoplanetary disks and their evolution. *Annu. Rev. Astron. Astrophys.* **49,** 67–117 (2011).
9. Barrado, Y., &. Navascués, D. On the age of the TW Hydrae association and 2M1207334–393254. *Astron. Astrophys.* **459,** 511–518 (2006).
10. Vacca, W. D. & Sandell, G. Near-infrared Spectroscopy of TW Hya: a revised spectral type and comparison with magnetospheric accretion models. *Astrophys. J.* **732,** 8 (2011).
11. Weintraub, D. A., Zuckerman, B. & Masson, C. R. Measurements of Keplerian rotation of the gas in the circumbinary disk around T Tauri. *Astrophys. J.* **344,** 915–924 (1989).
12. Calvet, N. *et al.* Evidence for a developing gap in a 10 Myr old protoplanetary disk. *Astrophys. J.* **568,** 1008–1016 (2002).
13. Thi, W.-F. *et al.* Herschel-PACS observation of the 10 Myr old T Tauri disk TW Hya. Constraining the disk gas mass. *Astron. Astrophys.* **518,** L125 (2010).
14. Gorti, U., Hollenbach, D., Najita, J. & Pascucci, I. Emission lines from the gas disk around TW Hydra and the origin of the inner hole. *Astrophys. J.* **735,** 90 (2011).
15. Natta, A. *et al.* in *Protostars and Planets V* 767–781 (Univ. Arizona Press, 2007).
16. Draine, B. T. *et al.* Dust masses, PAH abundances, and starlight intensities in the SINGS galaxy sample. *Astrophys. J.* **663,** 866–894 (2007).
17. Pilbratt, G. *et al.* The Herschel Space Observatory. *Astron. Astrophys.* **518,** 3–8 (2010).
18. Poglitsch, A. *et al.* The Photodetector Array Camera and Spectrometer (PACS) on the Herschel Space Observatory. *Astron. Astrophys.* **518,** L2 (2010).
19. Hoff, W., Henning, T. & Pfau, W. The nature of isolated T Tauri stars. *Astron. Astrophys.* **336,** 242–250 (1998).
20. Linsky, J. L. Deuterium abundance in the local ISM and possible spatial variations. *Space Sci. Rev.* **84,** 285–296 (1998).
21. Lissauer, J. J. & Stevenson, D. J. in *Protostars and Planets V* 591–606 (Univ. Arizona Press, 2007).
22. Greaves, J. S. & Rice, W. K. M. Have protoplanetary discs formed planets? *Mon. Not. R. Astron. Soc.* **407,** 1981–1988 (2010).
23. Mordasini, C., Alibert, Y., Benz, W., Klahr, H. & Henning, T. Extrasolar planet population synthesis. IV. Correlations with disk metallicity, mass, and lifetime. *Astron. Astrophys.* **541,** A97 (2012).
24. Tachihara, K., Neuhäuser, R. & Fukui, Y. Search for remnant clouds associated with the TW Hya association. *Publ. Astron. Soc. Jpn* **61,** 585–591 (2009).
25. Hogerheijde, M. R. *et al.* Detection of the water reservoir in a forming planetary system. *Science* **334,** 338–340 (2011).
26. Dartois, E., Dutrey, A. & Guilloteau, S. Structure of the DM Tau outer disk: probing the vertical kinetic temperature gradient. *Astron. Astrophys.* **399,** 773–787 (2003).
27. Wright, C. M., van Dishoeck, E. F., Cox, P., Sidher, S. D. & Kessler, M. F. *Infrared Space Observatory*–Long Wavelength Spectrometer detection of the 112 micron HD $J = 1 \rightarrow 0$ line toward the Orion Bar. *Astrophys. J.* **515,** L29–L33 (1999).
28. Bertoldi, F., Timmermann, R., Rosenthal, D., Drapatz, S. & Wright, C. M. Detection of HD in the Orion molecular outflow. *Astron. Astrophys.* **346,** 267–277 (1999).
29. Neufeld, D. A. *et al.* Spitzer observations of hydrogen deuteride. *Astrophys. J.* **647,** L33–L36 (2006).
30. Brinch, C. & Hogerheijde, M. R. LIME - a flexible, non-LTE line excitation and radiation transfer method for millimeter and far-infrared wavelengths. *Astron. Astrophys.* **523,** A25 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements Herschel is an ESA space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA. Support for this work was provided by NASA through an award issued by JPL/Caltech and by the US National Science Foundation under grant 1008800. This paper makes use of the following Atacama Large Millimeter/submillimeter Array (ALMA) data: ADS/JAO.ALMA#2011.0.00001.SV. ALMA is a partnership of ESO (representing its member states), the NSF (USA) and NINS (Japan), together with the NRC (Canada) and the NSC and ASIAA (Taiwan), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ.

Author Contributions E.A.B., L.I.C., U.G. and K.Z. performed the detailed calculations used in the analysis. J.D.G. reduced the Herschel data. S.M.A. provided detailed disk physical models and U.G. provided thermochemical models, both developed specifically for TW Hya. E.A.B. wrote the manuscript with revisions by N.J.E. All authors were participants in the discussion of results, determination of the conclusions and revision of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed E.A.B. (ebergin@umich.edu).

doi:10.1038/nature11733

Magnetic ratchet for three-dimensional spintronic memory and logic

Reinoud Lavrijsen[1], Ji-Hyun Lee[1], Amalio Fernández-Pacheco[1], Dorothée C. M. C. Petit[1], Rhodri Mansell[1] & Russell P. Cowburn[1]

One of the key challenges for future electronic memory and logic devices is finding viable ways of moving from today's two-dimensional structures, which hold data in an *x*–*y* mesh of cells, to three-dimensional structures in which data are stored in an *x*–*y*–*z* lattice of cells. This could allow a many-fold increase in performance. A suggested solution is the shift register[1,2]—a digital building block that passes data from cell to cell along a chain. In conventional digital microelectronics, two-dimensional shift registers are routinely constructed from a number of connected transistors. However, for three-dimensional devices the added process complexity and space needed for such transistors would largely cancel out the benefits of moving into the third dimension. 'Physical' shift registers, in which an intrinsic physical phenomenon is used to move data near-atomic distances, without requiring conventional transistors, are therefore much preferred. Here we demonstrate a way of implementing a spintronic unidirectional vertical shift register between perpendicularly magnetized ferromagnets of subnanometre thickness, similar to the layers used in non-volatile magnetic random-access memory[3]. By carefully controlling the thickness of each magnetic layer and the exchange coupling between the layers, we form a ratchet that allows information in the form of a sharp magnetic kink soliton to be unidirectionally pumped (or 'shifted') from one magnetic layer to another. This simple and efficient shift-register concept suggests a route to the creation of three-dimensional microchips for memory and logic applications.

The soliton is a magnetic frustration in a superlattice of perpendicularly magnetized ferromagnetic layers (referred to here as 'layers') coupled antiferromagnetically[4]. In Fig. 1a four different magnetic configurations of a six-layer superlattice are shown. The magnetic configuration of the ground state is a single-phase domain, which exists in two configurations as indicated. If the magnetization direction of the top three layers are flipped, a frustration is introduced where the two antiphase domains meet (indicated by an asterisk). These are sharp kink solitons (referred to here as 'solitons')[5–8]; they are sharp because they are only formed by two layers, and they are stable because all layers either above or below need to be switched to return to the ground state. The soliton has the interesting property that it can selectively be manipulated with an external magnetic field or potentially by a spin polarized current: for example, switching one of the layers forming the soliton moves it up or down. Furthermore, modern fabrication technology allows the functional layers to be of nearly atomic thickness.

In order to obtain field synchronized unidirectional propagation of solitons, we engineer the antiferromagnetic coupling J (in Oe nm) between the layers, and the thickness t (in nm) of the layers, in such a way that a ratchet scheme for unidirectional soliton propagation is inherent to the superlattice. (Note that for clarity we use $J > 0$ for antiferromagnetic coupling[9].) This is shown in Fig. 1b, where we alternate the coupling ($J_1 > J_2$) and the thicknesses ($t_1 < t_2$). As indicated by the green arrows, the lowest-energy position for a soliton to reside is between layers coupled by J_2, because $J_2 < J_1$. Assuming an Ising-macrospin approach, where the layers switch as a single domain (macrospin) and the magnetization is always aligned along the easy axis (Ising)[10], the switching field of each individual layer in the superlattice can be calculated as the sum of its coercive field H_c and the total exchange field J_{tot}/t due to its nearest neighbours (see Supplementary Information 1). For our discussion we use the parameters listed in Fig. 1d, which are taken from our experimental system (described

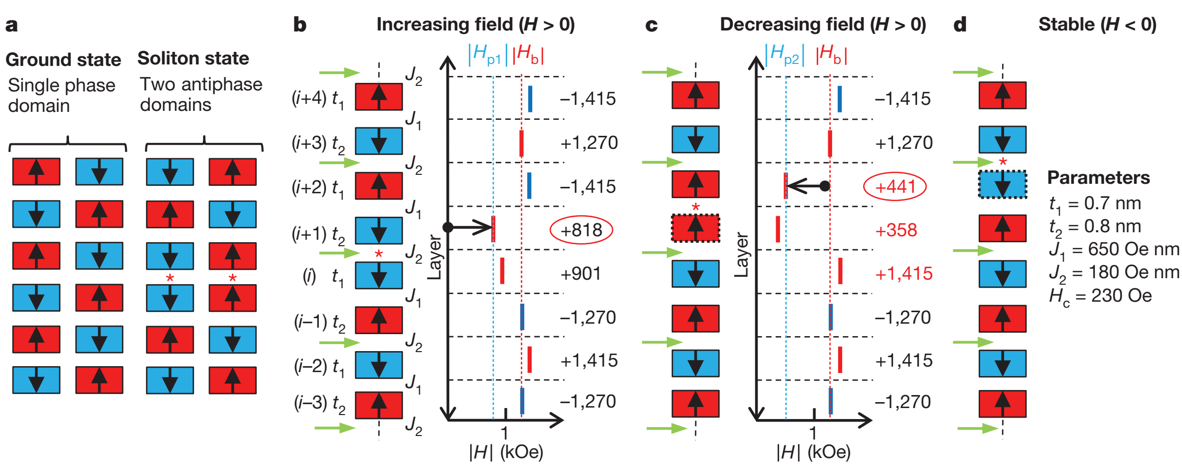


Figure 1 | Diagrams of solitons and ratchet scheme. a, Ground and soliton states of a six-layer superlattice. The blue and red boxes indicate magnetic layers magnetized down or up, respectively, as indicated by the black arrows. **b**, **c**, Schematic representation of an infinite superlattice with a propagating soliton. The labelling between parentheses (for example, $(i + 1)$) indicates the layer number, t_1 and t_2 indicate the layer thickness, J_1 and J_2 indicate the coupling between two layers. The green arrows indicate the lowest energy positions for a soliton. The panels consist of three parts: left, the superlattice configuration; middle, a graph indicating the switching field of every layer for the given configuration; right, the value of the switching field. **d**, Superlattice configuration after a full field cycle, showing the soliton propagated two layers upwards. In the list the parameters used to calculate the switching fields in **b** and **c** are given.

[1]Thin Film Magnetism Group, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK.

later). For example, for the configuration shown in Fig. 1b, where the soliton is formed by two layers coupled by J_2, layer $i+1$ will switch at $H_c + (J_1 - J_2)/t_2 = 818$ Oe, as it is anti-parallel to layer $i+2$ stabilizing it by J_1, but parallel to layer i destabilizing it by J_2. The switching field, calculated in this way for every layer, is shown to the right of the superlattice in Fig. 1b, where a red (blue) bar indicates if the switch will happen at positive (negative) applied field. As expected, the layers forming the soliton have the lowest switching field.

To propagate the soliton, we consider a square-wave field sequence with an amplitude H that is enough to switch layer $i+1$, indicated by propagation field 1 $H_{p1} = 818$ Oe, but lower than the bulk nucleation field, indicated by $|H_b| = 1{,}270$ Oe. H_b is defined as the field at which all t_2 layers aligned anti-parallel to the applied field switch, erasing the soliton from the superlattice. As the field steps to $+H$, layer $i+1$ switches first as it has the lowest switching field. This moves the soliton one layer up, leading to the new configuration shown in Fig. 1c (switched layer indicated by dotted outline). As layer $i+1$ has switched, its switching field and that of its two neighbouring layers have changed (values in red). This configuration is stable for $H_{p2} < H < H_b$ because the applied field is now parallel with the layers forming the soliton. Now the applied field steps back to zero, and the first switching field that is reached (decreasing field) is that of layer $i+2$ at propagation field 2: $H_{p2} = 441$ Oe. This moves the soliton one layer further up and the configuration shown in Fig. 1d is obtained at remanence. The switching fields are now identical to those in Fig. 1b, but with the soliton two layers up. In the second half of the field cycle (negative field step), the magnetic configuration shown in Fig. 1d is preserved, as H is aligned parallel to the layers forming the soliton and $|H| < |H_b|$. Note that if we had started with a soliton with the opposite polarity, that is, layers forming the soliton pointing upwards in Fig. 1b, it would propagate on the negative part of a field cycle and be stationary on the positive part. This demonstrates the essence of the soliton ratchet propagation scheme, which is made intrinsic to the superlattice by the alternating coupling between the layers and their alternating thicknesses. Interestingly, the arguments just given can be generalized to the case where several solitons can be propagated up the superlattice using the same field sequence, thus implementing a full serial shift register.

By analysing the ratchet soliton propagation as presented in Fig. 1, three operating limits can be deduced. The first can be regarded as a 'field-pressure' for unidirectional soliton propagation, which is given by the difference between the switching field values of the layers propagating the soliton up or down. This is given by $(J_1 - J_2)(t_1 - t_2)/(t_1 t_2) = -84$ Oe, which is the difference between the switching field of layers $i+1$ and i in Fig. 1b and layers $i+2$ and $i+1$ in Fig. 1c. When this field pressure is negative (positive) the superlattice will act as a ratchet for upward (downward) propagation. Hence, the conditions set before, $J_1 > J_2$, $t_1 < t_2$, ensure upward soliton ratchet action. The second limit is $|H_b| - |H_{p1}| > 0$ giving $2J_2/t_2 > 0$ (450 Oe for the parameters used above); this ensures that H_{p1} is separated from the bulk switching field H_b. The third limit is $|H_{p1}| - |H_{p2}| > 0$, giving $2H_c > (J_1 - J_2)(t_2 - t_1)/(t_1 t_2)$ ($460 > 84$ for the parameters used above), which ensures that a soliton propagates one layer per field step, that is, it is not directly expelled out of the superlattice when the applied field reaches H_{p1}.

We now present experimental proof of controlled injection and synchronous vertical propagation of a soliton through a sputtered antiferromagnetically coupled superlattice, as described above. The superlattice studied here, with 44 magnetic and non-magnetic layers in total, of which 11 are magnetic layers (referred to in the text as 'layers'), is schematically shown in Fig. 2a (see Methods). It consists of two regions; the first three layers (labelled M1 to M3) act as a soliton injector, and the top eight layers (M4 to M11) act as the soliton propagation region. The injector consists of three thin layers of identical thickness (t_0); these layers are highly antiferromagnetically-coupled compared to the layers in the propagation region (that is, $J_0 > J_1, J_2$). The soliton propagation region follows the superlattice sequence discussed above for upward soliton propagation.

To explain soliton injection in the superlattice, we present in Fig. 2b the normalized hysteresis curve obtained using a vibrating sample magnetometer (VSM). The solid line is a fit to this curve using the Ising-macrospin model with parameters as given in the figure legend. A good agreement is found except for the transition labelled with an exclamation mark (!), which is discussed in Supplementary Information 2. In Fig. 2c, the magnetic configuration of the superlattice going from negative saturation to remanence (configurations 1–4, as shown in the red filled circles at the bottom of the panel) is shown, as inferred from the Ising-macrospin model. As the field reduces from negative saturation (configuration 1), the first layer to switch is M2 as it is thin and highly antiferromagnetically coupled to M1 and M3, leading to configuration 2. Now M1 and M3 are stabilized by M2 as they are aligned anti-parallel. Using the same reasoning, layers M5, M7 and M9 will switch next (configuration 3). Just before reaching remanence M11 switches (configuration 4) as it is antiferromagnetically-coupled only to a single neighbour (M10). At remanence all layers are now oriented anti-parallel to their neighbour(s) except for layers M3 and M4, which are both pointing downwards. Hence, by saturating and relaxing to remanence a soliton with negative polarity pointing in the direction of the original saturating field is injected.

In Fig. 2e we present a VSM measurement of a soliton injected and propagated through the superlattice as a function of cycle number, using a square-wave field sequence with amplitude $H = 850$ Oe (Fig. 2d). We start by injecting a soliton between M3 and M4, as explained above (in Fig. 2c, configurations 1–4, indicated by numbers in red filled circles in that panel and in Fig. 2e). The magnetic configuration of the superlattice under the applied field sequence is traced schematically in Fig. 2c. On the first (positive) field step, the first layer to switch is M4 (the same as transition H_{p1} in Fig. 2b). When the field steps back to zero, M5 switches and the soliton has moved two layers up. This is identified by the increase in magnetic moment between configurations 4 and 6. Because the two antiphase domains have a different total magnetic moment, the expansion of the lower antiphase domain at the cost of the top one gives an increase in normalized magnetic moment of $2(t_2 - t_1)/t_{tot} \approx 0.026$, as observed. Continuing on the negative field step of the cycle, we observe no change in the magnetic configuration as expected. In the following cycles the soliton follows the same routine (configurations 6–11), and is expelled out of the superlattice after a total of four field cycles (eight layers, two layers per cycle) leading to an empty (configuration E) superlattice. On further field cycles, no switches or changes in moment are observed, as expected. Two cycles after the soliton was expelled from the superlattice, we inject an inverted soliton (soliton with positive polarity) by saturating the superlattice in a positive field (configuration 1, blue filled circle). This soliton follows the same routine, with the difference that it now propagates on the negative field steps. The blue filled circles refer to the inverted superlattice configuration of Fig. 2c. This shows the experimental realization of the soliton ratchet scheme, corresponding exactly with the behaviour expected from a simple Ising-macrospin model.

To complement the quasi-static VSM measurements, we present in Fig. 3 the propagation of a soliton using measurements of the magneto-optical Kerr effect (MOKE) and a pulse coil. The field sequence applied to the superlattice is shown in Fig. 3a, where ~250-μs single sine wave field pulses, as shown in the inset above the graph, are applied with 5-ms spacing. In Fig. 3b the associated averaged MOKE signal is shown. We start with an empty superlattice and inject a soliton pointing down at time 0 ms by applying a large field pulse, which is shown by the drop in MOKE signal after the injection field pulse. The MOKE level depends on the depth of the soliton in the superlattice owing to the skin depth of the laser light. On subsequent field pulses with lower amplitude ($H_{p1} < H < H_b$), the soliton is propagated up the superlattice, as observed by the changing MOKE levels.

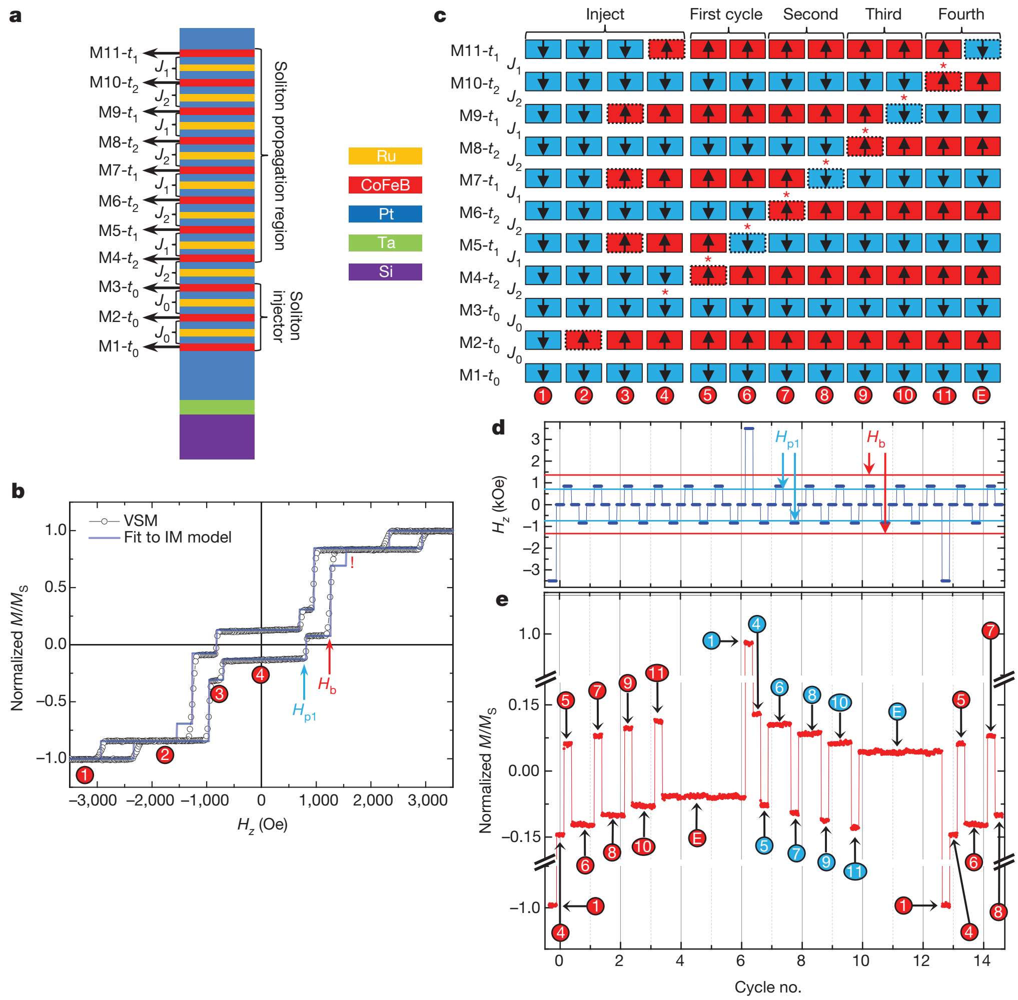


Figure 2 | Superlattice stack sequence, major hysteresis loop and soliton propagation. **a**, Schematic illustration of the experimental superlattice. M_1–M_{11} indicate the magnetic layers in the superlattice, t_0, t_1 and t_2 indicate the layer thickness, and J_0, J_1 and J_2 indicate the coupling between two layers. The colour coding of the superlattice indicating the materials is shown to the right of the superlattice. **b**, Normalized (M/M_S) VSM major hysteresis loop as a function of applied field H_z with fit to Ising-macrospin model using the following parameters: $t_0 = 0.6$ nm, $t_1 = 0.7$ nm, $t_2 = 0.8$ nm, $J_0 = 790$ Oe nm, $J_1 = 650$ Oe nm, $J_2 = 180$ Oe nm and $H_c = 230$ Oe. The red and blue arrows indicate H_b and H_{p1}, respectively. The numbers in red filled circles correspond to the superlattice configuration shown in **c**. **c**, Magnetic configuration of the superlattice at every level indicated in **b** and **e**. **d**, Square-field cycle. The red and blue arrows indicate H_b and H_{p1}, respectively. **e**, Normalized (M/M_S) VSM signal obtained when the square-field cycle as shown in **d** is applied. The numbers in blue filled circles indicate the same configuration as in **c**, albeit with the orientation of every layer inverted.

The superlattice configuration at each level is again indicated by the red filled circles, corresponding to the superlattice configurations in Fig. 2c. This shows that it follows the same propagation routine as before. The spikes and ringing in the MOKE signal during and directly after the propagation field pulses are partly due to an induction effect of the large field gradients present ($>$2,000 T s^{-1} during the low amplitude propagation pulses). However, the increasing MOKE signal amplitude also shows the propagation of the soliton towards the top of the superlattice, as indicated by the dashed blue line in Fig. 3b. In Fig. 3c–e, we show the propagation of a soliton when we apply a pulse train of 2, 3 and 4 pulses at time 5 ms. The levels observed after the pulse trains correspond to the levels in Fig. 3b, confirming field-synchronized propagation of a soliton with a fast pulse train. The minimum field required for propagation, H_{p1}, has increased from 818 Oe in the quasi-static VSM measurements to ~1,100 Oe, owing to thermally induced processes increasing the H_c values of all the layers at high field-sweep rates[11–13]. We do not believe this is a problem, as the perpendicularly magnetized CoFeB used here is widely used in high speed magnetic random access memory (MRAM) devices[3,14,15].

Our experiments show how a stable and sharp magnetic kink soliton can be injected and propagated vertically through a superlattice of antiferromagnetically coupled ultrathin layers. We obtain the equivalent electronic functionality of ~20 transistors within a vertical length of only 2 nm, showing true atomic-scale digital logic operation[16]. Furthermore, the possibility of selectively propagating a negative or positive soliton by field polarity could allow complex logic operations to be performed within a data shift register. We visualize a superlattice with hundreds of layers, where solitons are injected at the bottom or top of the superlattice and unidirectionally pumped to the other side. This would boost the data-storage density of conventional MRAM device architecture[15], as multiple bits could be stored in each cell using the same number of transistors. The writing and reading of the solitons could be carried out in the same way as already used in conventional MRAM technology[3]. For example, by adding a magnetic tunnel junction at the two extremes of the superlattice, one optimized for reading solitons as they leave the superlattice by the tunnel magnetoresistance effect, and the other optimized for soliton injection into the superlattice by spin transfer torque. Because we have opened up the third dimension, the pressure to continually shrink the lateral size of devices is reduced (see Supplementary Information 5). However, further quantitative investigation is needed to address fundamental technical challenges for the proposed soliton-based device concept at laterally reduced dimensions, such as dispersion in properties[17], dipole fields[18–20], data retention lifetime[15] and the reading and writing

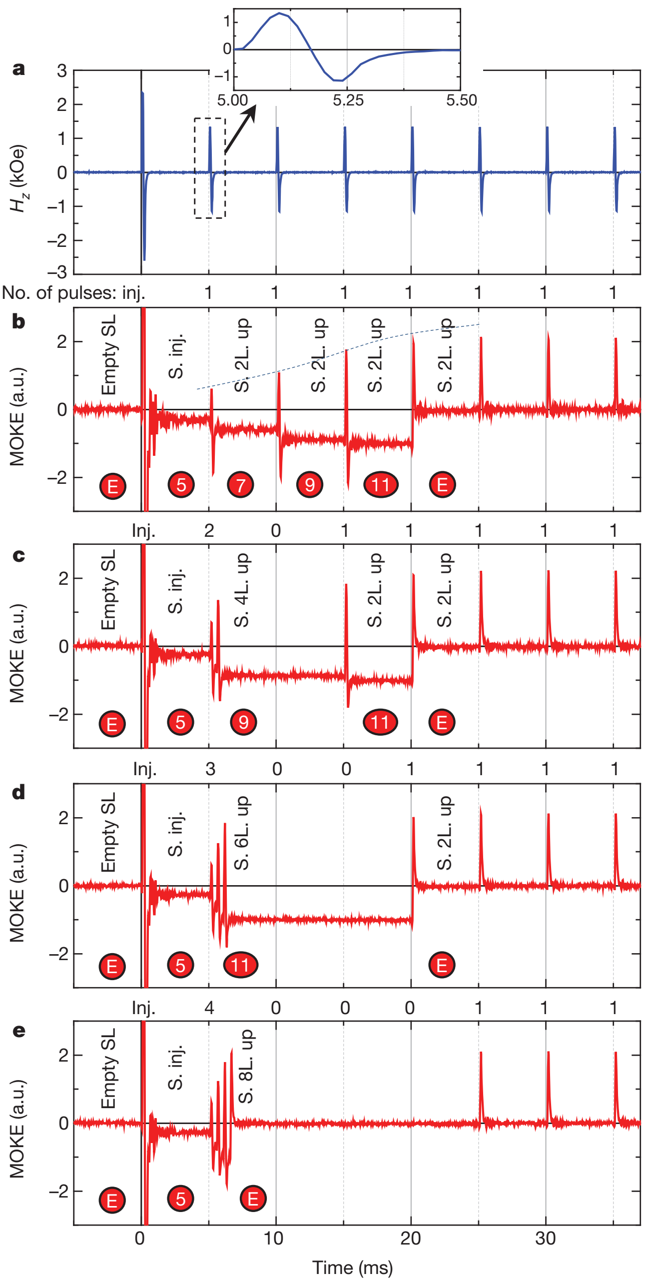


Figure 3 | Soliton propagation with field pulses. **a**, Sine-wave field pulse sequence as a function of time. The inset shows a single field pulse. **b**, Averaged MOKE (magneto-optical Kerr effect) signal (600 averages) recorded during the field cycle as shown in **a** with a single pulse every 5 ms as indicated above the graph. The numbers in red filled circles correspond to the configurations shown in Fig. 2c. The blue dotted line indicates the increasing MOKE intensity as the soliton propagates towards the top of the superlattice. **c**–**e**, MOKE signal as a function of time showing a soliton propagated with multiple consecutive pulses as indicated above the graphs. Empty SL refers to an empty superlattice as given in configuration E in Fig. 2c. S. Inj. refers to soliton injected, S. 2L up refers to the soliton propagated two layers up, and so on.

of solitons. These challenges are further discussed in Supplementary Information 3–5. Our present work paves the way to shift spintronics into the third dimension.

METHODS SUMMARY

Samples are grown using d.c. magnetron sputtering on Si/SiO$_x$ substrates, Ta(2)/Pt(20) buffer (thickness in nm) and capped with 2 nm Pt to prevent oxidation. The CoFeB composition is $Co_{60}Fe_{20}B_{20}$ (in atom %). The base pressure of the sputter chamber is $\sim 3 \times 10^{-8}$ mbar. The CoFeB layers are antiferromagnetically coupled to each other by a 0.9-nm Ru spacer via the Ruderman-Kittel-Kasuya-Yosida (RKKY) mechanism. The coupling strength (J_0, J_1, J_2) between the layers is set by inserting different Pt thicknesses between the Ru and the CoFeB; inserted Pt thickness for J_0, J_1, J_2 is 0.46, 0.52, 0.72 nm, respectively. We use vibrating sample magnetometery, magneto-optical Kerr effect, and the anomalous Hall effect to determine the magnetic properties. The external field is applied perpendicular to the sample plane unless specified otherwise. For the MOKE measurements, we use custom pulse coils with an L/R time constant of 0.3 ms. The current pulses required are driven by a gradient amplifier controlled by an arbitrary function generator allowing for synchronized triggering of the pulses and data acquisition by an oscilloscope.

Received 19 June; accepted 31 October 2012.

1. Parkin, S. S. P., Hayashi, M. & Thomas, L. Magnetic domain-wall racetrack memory. *Science* **320,** 190–194 (2008).
2. Allwood, D. A. *et al.* Magnetic domain-wall logic. *Science* **309,** 1688–1692 (2005).
3. Kawahara, T., Ito, K., Takemura, R. & Ohno, H. Spin-transfer torque RAM technology: review and prospect. *Microelectron. Reliab.* **52,** 613–627 (2012).
4. Hellwig, O., Berger, A., Kortright, J. B. & Fullerton, E. E. Domain structure and magnetization reversal of antiferromagnetically coupled perpendicular films. *J. Magn. Magn. Mater.* **319,** 13–55 (2007).
5. Wang, R. W., Mills, D. L., Fullerton, E. E., Mattson, J. E. & Bader, S. D. Surface spin-flop transition in Fe/Cr(211) superlattices—experiment and theory. *Phys. Rev. Lett.* **72,** 920–923 (1994).
6. Mühlbauer, S. *et al.* Skyrmion lattice in a chiral magnet. *Science* **323,** 915–919 (2009).
7. Seki, S., Yu, X. Z., Ishiwata, S. & Tokura, Y. Observation of skyrmions in a multiferroic material. *Science* **336,** 198–201 (2012).
8. Baryakhtar, V. G., Chetkin, M. V., Ivanov, B. A. & Gadetskii, S. N. *Dynamics of Topological Magnetic Solitons: Experiments and Theory* (Springer, 1994).
9. Lavrijsen, R. *et al.* Tuning the interlayer exchange coupling between single perpendicularly magnetized CoFeB layers. *Appl. Phys. Lett.* **100,** 052411 (2012).
10. Kronmller, H. & Parkin, S. (eds) *Handbook of Magnetism and Advanced Magnetic Materials* Vols 1–5 (Wiley, 2007).
11. Lavrijsen, R. *et al.* Reduced domain wall pinning in ultrathin Pt/Co$_{100-x}$B$_x$/Pt with perpendicular magnetic anisotropy. *Appl. Phys. Lett.* **96,** 022501 (2010).
12. Lemerle, S. *et al.* Domain wall creep in an Ising ultrathin magnetic film. *Phys. Rev. Lett.* **80,** 849–852 (1998).
13. Bruno, P. *et al.* Hysteresis properties of ultrathin ferromagnetic films. *J. Appl. Phys.* **68,** 5759–5766 (1990).
14. Sbiaa, R., Meng, H. & Piramanayagam, S. N. Materials with perpendicular magnetic anisotropy for magnetic random access memory. *Phys. Status Solidi RRL* **5,** 413–419 (2011).
15. Gajek, M. *et al.* Spin torque switching of 20nm magnetic tunnel junctions with perpendicular anisotropy. *Appl. Phys. Lett.* **100,** 132408 (2012).
16. Langholz, G., Kandel, A. & Mott, J. L. *Foundations of Digital Logic Design* (World Scientific, 1998).
17. Thomson, T., Hu, G. & Terris, B. D. Intrinsic distribution of magnetic anisotropy in thin films probed by patterned nanostructures. *Phys. Rev. Lett.* **96,** 257204 (2006).
18. Baltz, V. *et al.* Multilevel magnetic nanodot arrays with out of plane anisotropy: the role of intra-dot magnetostatic coupling. *Eur. Phys. J. Appl. Phys.* **39,** 33–38 (2007).
19. Baltz, V. *et al.* Balancing interlayer dipolar interactions in multilevel patterned media with out-of-plane magnetic anisotropy. *Appl. Phys. Lett.* **94,** 052503 (2009).
20. Tudosa, I., Katine, J. A., Mangin, S. & Fullerton, E. E. Perpendicular spin-torque switching with a synthetic antiferromagnetic reference layer. *Appl. Phys. Lett.* **96,** 212504 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements R.L. was supported by the Netherlands Organization for Scientific Research and Marie Curie Cofund Action (NWO-Rubicon 680-50-1024). A.F.-P. was supported by a Marie Curie IEF within the Seventh European Community Framework Programme No. 251698; 3DMAG-NANOW. We acknowledge research funding from the European Community under the Seventh Framework Programme Contract No. 247368: 3SPIN.

Author Contributions R.L. and R.P.C. planned the experiment; R.L. fabricated the samples; R.L. and J.-H.L. performed the experiments; D.C.M.C.P. performed the dipole field calculations; R.L. analysed the data and wrote the manuscript. All authors discussed the results and contributed to the scientific interpretation as well as to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.P.C. (rpc12@cam.ac.uk).

doi:10.1038/nature11839

Responsive biomimetic networks from polyisocyanopeptide hydrogels

Paul H. J. Kouwer[1]*, Matthieu Koepf[1]*, Vincent A. A. Le Sage[1], Maarten Jaspers[1], Arend M. van Buul[1], Zaskia H. Eksteen-Akeroyd[1], Tim Woltinge[1], Erik Schwartz[1], Heather J. Kitto[1], Richard Hoogenboom[1]†, Stephen J. Picken[2], Roeland J. M. Nolte[1], Eduardo Mendes[2] & Alan E. Rowan[1]

Mechanical responsiveness is essential to all biological systems down to the level of tissues and cells[1,2]. The intra- and extracellular mechanics of such systems are governed by a series of proteins, such as microtubules, actin, intermediate filaments and collagen[3,4]. As a general design motif, these proteins self-assemble into helical structures and superstructures that differ in diameter and persistence length to cover the full mechanical spectrum[1]. Gels of cytoskeletal proteins display particular mechanical responses (stress stiffening) that until now have been absent in synthetic polymeric and low-molar-mass gels. Here we present synthetic gels that mimic in nearly all aspects gels prepared from intermediate filaments. They are prepared from polyisocyanopeptides[5–7] grafted with oligo(ethylene glycol) side chains. These responsive polymers possess a stiff and helical architecture, and show a tunable thermal transition where the chains bundle together to generate transparent gels at extremely low concentrations. Using characterization techniques operating at different length scales (for example, macroscopic rheology, atomic force microscopy and molecular force spectroscopy) combined with an appropriate theoretical network model[8–10], we establish the hierarchical relationship between the bulk mechanical properties and the single-molecule parameters. Our results show that to develop artificial cytoskeletal or extracellular matrix mimics, the essential design parameters are not only the molecular stiffness, but also the extent of bundling. In contrast to the peptidic materials, our polyisocyanide polymers are readily modified, giving a starting point for functional biomimetic hydrogels with potentially a wide variety of applications[11–14], in particular in the biomedical field.

The artificial gels are based on polyisocyanopeptides (PICs), composed of a β-helical architecture stabilized by a peptidic hydrogen-bond network along the polymer backbone[6]. Polymers **P1**–**P3** were obtained through a nickel(II)-catalysed polymerization of di-, tri- and tetraethylene glycol functionalized isocyano-(D)-alanyl-(L)-alanines **1–3** (Fig. 1)[15]. Variation of the catalyst to monomer ratio allowed us to tune the molecular weights of the polymers (see Supplementary Information). The hydrogen-bonded helical structure of the polymer

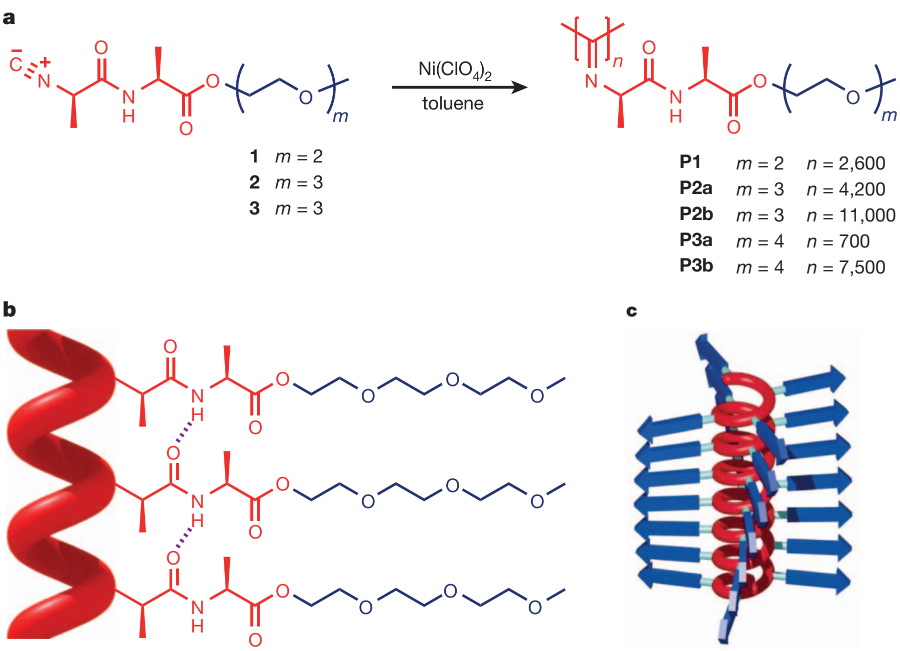


Figure 1 | Oligo(ethylene glycol)-substituted PICs. a, Synthesis of the polymers—the degree of polymerization is estimated from atomic force microscopy (AFM) experiments. **b**, Representation of the hydrogen-bond network (dotted lines) that stabilizes the secondary helical structure for **P2**. **c**, Schematic illustration of the 4_1 β-sheet helix. Colour coding: red, the stiff helical polyisocyanide backbone stabilized with the hydrogen-bonded dialanyl groups; in **b** and **c** the backbone is schematically shown as a helix. Blue, the ethylene glycol peptide substituent 'tails', represented in **c** by blue arrows. Panel **c** is from ref. 6, reprinted with permission from AAAS.

[1]Radboud University Nijmegen, Institute for Molecules and Materials, Department of Molecular Materials, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands. [2]Delft University of Technology, Department of NanoStructured Materials, Julianalaan 136, 2628 BL Delft, The Netherlands. †Present address: Supramolecular Chemistry Group, Department of Organic Chemistry, Ghent University, Krijgslaan 281-S4, 9000 Ghent, Belgium.
*These authors contributed equally to this work.

backbone was confirmed by infrared and circular dichroism (Supplementary Fig. 1) spectroscopies. In aqueous solution and in the gel phase, the secondary structure of the polymer is stable up to about 70 °C as shown with circular dichroism experiments (Supplementary Figs 2 and 3). The combination of a densely packed helical structure and the strong intramolecular hydrogen bonds gives stiff polymer chains that are readily visualized by atomic force microscopy (AFM; Fig. 2a and Supplementary Fig. 4).

Thermal analysis of dilute aqueous solutions of **P2b** and **P3b** showed the formation of transparent hydrogels on heating at 18 and 44 °C, respectively. Polymers **P2a** and **P3a** did not form gels, but precipitated at these temperatures forming a cloudy suspension, whereas **P1** failed to dissolve in water (we attribute this to a transition temperature below 0 °C). The sol–gel phase transition was very fast (on a timescale of seconds) and fully reversible (Supplementary Fig. 13). The structure of the gel was visualized by AFM (Fig. 2b and Supplementary Fig. 5a–f) and cryo scanning electron microscopy (SEM; Supplementary Fig. 5g). Both techniques showed a network composed of bundles of polymer chains. The extent of bundling was estimated by statistical analysis of the AFM images of the bundles and the isolated polymer chains (Fig. 2c). The narrow distributions of relative heights was used to abstract the bundle number (average number of polymer chains per bundle) $N = 6.9$, using the relation $N = d_B^2/d_0^2 \approx h_B^2/h_0^2$, in which d_0 and d_B are the diameters, and h_0 and h_B the heights, of isolated chains and bundles, respectively.

We found that the bundle dimensions were constant irrespective of the polymer concentration. AFM analysis (Supplementary Fig. 6) of samples at higher concentration shows more rather than thicker bundles, which is also indicated by preliminary single-particle tracking studies of gels of **P2b** that show nanoparticle diffusion coefficients that strongly scale with concentration (Supplementary Fig. 7). The latter confirms that at higher concentrations more bundles (and hence smaller pores, which result in restricted particle displacement) are formed. This self-limiting behaviour of bundle formation is thought to be related to the chiral nature (that is, the helicity) and the intrinsic stiffness of the polymer molecules[16]. As a consequence of the fixed bundle size, the average pore size in the gel is directly controlled by the polymer concentration. Chain bundling is commonly observed for cytoskeletal polymers and the bundle properties (dimensions,

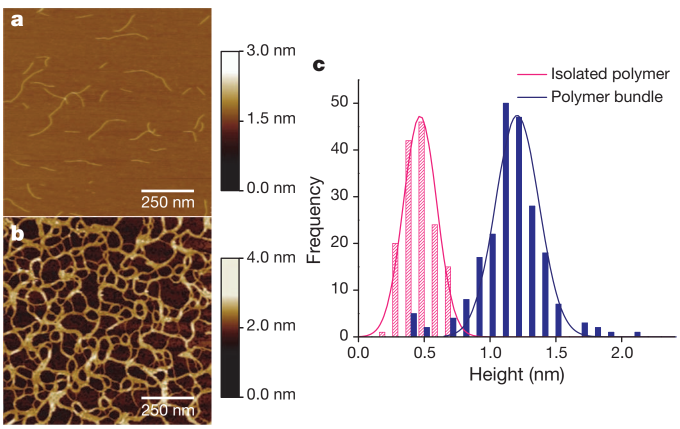


Figure 2 | AFM analysis of polymers and gel. a, AFM image of isolated polymer chains of **P2b**, spin-coated from an organic solution on mica. Colour scale here and in **b** shows height. **b**, AFM image of a 'monolayer' of bundles of the **P2b** gel transferred to mica. Occasional non-bundled polymers are visible. **c**, Statistical height histograms of both isolated chains (pink) and bundles (blue). Both show similarly narrow Gaussian distributions (see fits) with chain height $h_0 = 0.46 \pm 0.13$ nm and bundle height $h_B = 1.2 \pm 0.2$ nm. We note that the absolute height found by AFM is consistently too low. Considering that the diameter of the peptidic polymer without the ethylene glycol substituents is roughly 2 nm, only relative height distributions can be used to estimate the bundle numbers.

stiffness) are critical parameters in the mechanical properties of those gels. For gels based on actin or intermediate filaments, bundling is controlled by additives, ranging from binding proteins[17] to divalent metal ions[18], whereas bundle formation in the PIC gels is thermally activated.

The process of thermally induced gel formation is attributed to hydrophobic effects caused by the ethylene glycol tails grafted onto the polyisocyanide backbone. Flexible oligo(ethylene glycol) grafted polymers have been reported to show sharp order–disorder phase transitions at the lower critical solution temperature (LCST)[19]. Previous studies have demonstrated a linear relationship between the transition temperature and the average length of the ethylene glycol tail over a broad temperature range[20]. Heating **P2** and **P3** results in the entropic desolvation of the ethylene glycol arms, giving rise to more hydrophobic chains that separate from the aqueous solution. Indeed, the low molar mass polymers **P2a** and **P3a** precipitate at the LCST, whereas longer polymers yield completely transparent gels at the transition, as the long chains are kinetically trapped in a network structure before they precipitate. Even at very low concentrations, the gels are able to support their own weight during vial inversion tests; a sample of **P2b** passed the test at concentrations as low as 0.006 wt%, (Supplementary Fig. 8), which is about an order of magnitude lower in concentration than many of the well-known synthetic superhydrogelators[21].

To learn more about their mechanical properties, the polymer gels were subjected to a full variable temperature rheological analysis. Samples were measured in a Couette configuration with small oscillatory deformations at different frequencies and amplitudes in the linear response regime (Supplementary Figs 9 and 10). A broad-range frequency sweep in the gel phase (Supplementary Fig. 10) corroborates that the crosslinks formed at the LCST are permanent in nature. Temperature sweeps of **P2b** and **P3b** (Fig. 3a and Supplementary Fig. 11) show, at low temperatures, liquid-like behaviour with a storage modulus G' lower than the loss modulus G''. The temperature of the sharp transition that marks gel formation depends on the length of the ethylene glycol tail. It shows little dependence on the polymer concentration c. At elevated temperatures G' levels off to a plateau G_0; its absolute value, however, scales strongly with c. Analysis showed a power-law behaviour, $G_0 \propto c^n$ with coefficients n of 2.2 and 2.7 for **P2b** and **P3b**, respectively. These experimental values are in line with the theory of permanently linked semi-flexible networks that display purely entropic elasticity[9] (in which $n = 11/5$), and with other experimental studies based on cross-linked cytoskeletal proteins like actin[10] and intermediate-filament gels[8] (with $n = 2-2.5$), and also with other stiff materials such as DNA gels ($n = 2.3$)[22].

Unlike many gels of synthetic polymers, biopolymer gels show a strong, and well-defined, nonlinear stress response after a critical stress σ_c has been applied[4] (stress-stiffening). Although its origins are currently being debated[4,23,24], the effect is well described experimentally[4,8,9]. In the nonlinear regime, a small increase in the strain γ gives very high stress levels and often results in the rupture of the gel. To probe this regime carefully, we used a recently benchmarked pre-stress protocol[25] and determined the differential modulus K (the real part of which is defined as $K' = d\sigma/d\gamma$) as a function of applied stress σ (Fig. 3b). When scaled to G_0 and σ_c (Fig. 3c) all curves of **P2b** at different concentrations and temperatures reduce to a single master curve, displaying the theoretically predicted $K' \propto \sigma^{3/2}$ dependency[8]. The PIC-based gels show a quantitative resemblance to the protein-based biogels, even in the nonlinear regime.

A theoretical model for semi-flexible networks, based on the extensible worm-like chain model[26], has been developed to explain the mechanical behaviour of actin[9] and intermediate-filament-based hydrogels[9]. This model considers the network as a collection of thermally fluctuating bundles, with l_c as the average length between the crosslinks between bundles. We have modified the existing network model[23] to quantitatively describe the unusual experimentally observed thermal behaviour and to account for the fact that the bundle size in our system

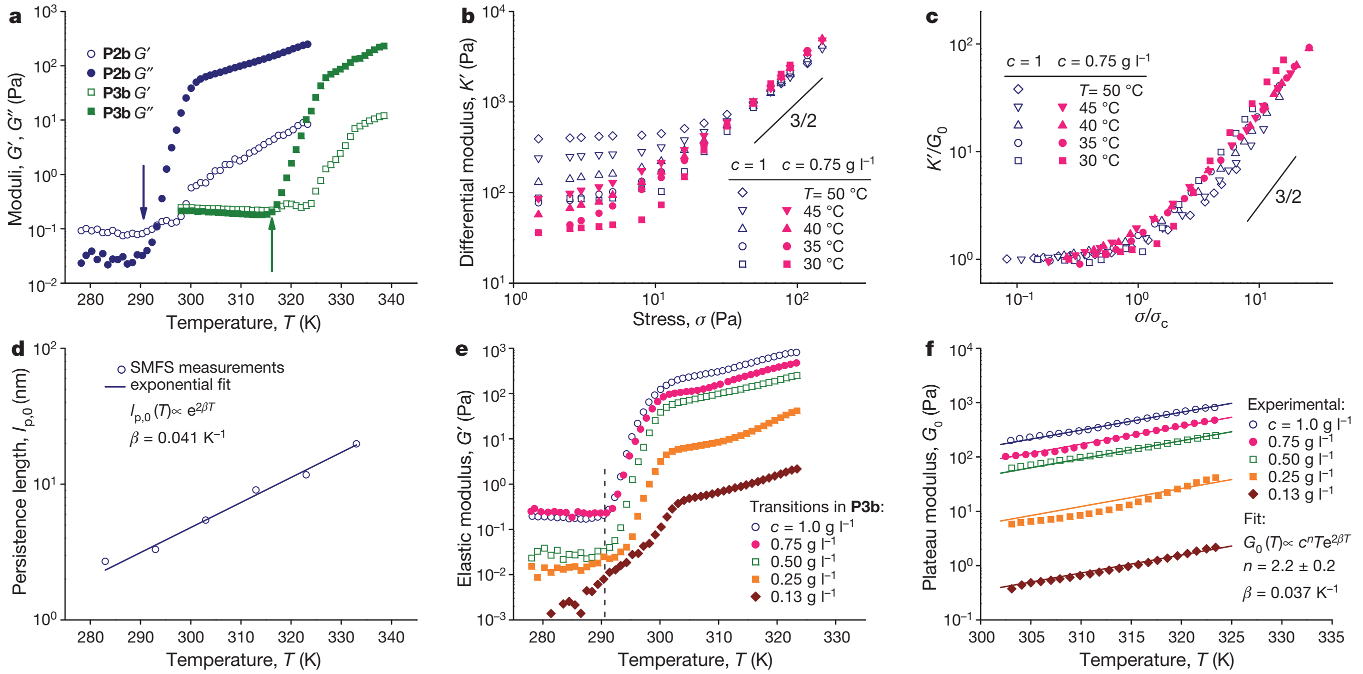


Figure 3 | Rheological analysis of PIC gels. **a**, Moduli G' and G'' as a function of temperature T for **P2b** and **P3b** at $c = 1.0\ \mathrm{mg\ ml^{-1}}$. The arrows indicate the transition temperature, rheologically determined as the onset of the step in G' at $\omega = 6.2\ \mathrm{rad\ s^{-1}}$ ($f = 1$ Hz, Supplementary Fig. 12). **b**, Differential modulus K' as a function of stress σ for different values of c and T. The model prediction $K' \propto \sigma^{3/2}$ is shown in **b** and **c**. **c**, Data scaled with the plateau modulus G_0 and the critical stress σ_c; all curves, independent of variations in c and T, collapse to a single master curve. **d**, Single chain persistence length $l_{p,0}$ as a function of T of **P2b** between 10 and 60 °C measured by SMFS, fitted to a single exponential as shown. **e**, G' as a function of T for **P2b** at different concentrations. The dashed line at $T = 18$ °C shows that the onset of the gel temperature is nearly concentration-independent. **f**, G_0 as a function of T and exponential fits to n and β for different concentrations.

is independent of the concentration. The details of the model are given in the Supplementary Information. Not only does this model describe our experimental results accurately, it also yields information about the critical microscopic parameters—such as the persistence length of the bundles, $l_{p,B}$, and l_c. To extract this information, we apply equations (1) and (2) to the experimentally determined macroscopic quantities G_0 and σ_c:

$$G_0 = 6\chi \frac{c}{N} RT \frac{l_{p,B}^2}{l_c^3} \tag{1}$$

$$\sigma_c = \chi \frac{c}{N} RT \frac{l_{p,B}}{l_c^2} \tag{2}$$

Here, χ combines molecular constants, R is the gas constant and T is the absolute temperature. Equations (1) and (2) show that G_0 and σ_c depend on N, $l_{p,B}$ and l_c (l_c in turn also depends on concentration). Rheological measurements in the linear and nonlinear regimes, with c and T as experimental variables, in combination with variable temperature single molecule force spectroscopy (SMFS) measurements, allowed us to calculate $l_{p,B}$, l_c and N.

SMFM measurements[27] provide information on the persistence length of individual polymer chains, $l_{p,0}$. In these experiments we determined force–distance curves of dilute polymer samples and subjected the results to the same extensible worm-like chain model that was applied to analyse the rheological data. Subsequent statistical analysis of the experimental data provided the average values for $l_{p,0}$ (Supplementary Fig. 14). SMFS measurements on **P2b**, equilibrated in water, typically yielded modest values for $l_{p,0}$ (Fig. 3d)[27], which is attributed to water weakening the hydrogen-bond network along the polymer backbone (see Supplementary Information). A temperature sweep between 10 and 60 °C showed an exponential increase of the persistence length $l_{p,0}(T) \propto e^{\beta T}$ with an exponent β of $0.041\ \mathrm{K^{-1}}$.

Figure 3e shows the plateau modulus $G_0(c,T)$ of **P2b** as obtained by bulk rheological temperature sweeps. At all c, G_0 showed an exponential increase with T (Fig. 3f); for this temperature range, the data were successfully fitted to $G_0(c,T) \propto c^n T e^{2\beta T}$ with exponents $n = 2.2$ and $2\beta = 0.073\ \mathrm{K^{-1}}$. In our system, the only temperature dependent contribution to G_0 is $l_{p,B}^2$ (equation (1)). The close match of the observed exponent from SFMS measurements and that from bulk rheology clearly indicates that the thermally induced increase in G_0 is simply the result of the stiffening of the individual polymer chains. This was confirmed by independent measurements of $\sigma_c(T)$ at different concentrations, which yielded a similar exponent, $\beta = 0.049\ \mathrm{K^{-1}}$ (equation (2)).

Combining equations (1) and (2) returns $l_{p,B}$ as a function of N, G_0 and σ_c; the last two values were experimentally determined by bulk rheology in the linear and nonlinear regime. By taking $N \approx 7$, as obtained from AFM measurements, a value of $l_{p,B}$ of the order of hundreds of nanometres for **P2b** was found, about two orders of magnitude larger than $l_{p,0}$. This difference can only be rationalized by considering that the chains in the bundles are strongly interacting, and behave effectively as a single fibre with the constituent polymer chains ‘glued’ together. This so-called tight bundle regime is characterized by a square dependence of $l_{p,B}$ with N: $l_{p,B} = l_{p,0}N^2$; this is in contrast to the loose bundle regime, which shows a linear relationship[28]. Cross-linked biofibres, such as actin, show a transition from the tight to the loose bundle regime with increasing N. In line with these results, we also find a square dependence at low bundle numbers. By establishing the regime in which the bundles interact, we can now calculate N by the straightforward comparison of the SMFS results and the (nonlinear) rheology data. Under the standard conditions ($1\ \mathrm{mg\ ml^{-1}}$, 30 °C), we find $N = 9.1$, which agrees closely with the value estimated from the AFM measurements. Calculations of N at different temperatures and concentrations yield very consistent numbers, further highlighting that, for our materials, the bundle characteristics are

Table 1 | Comparison of hydrogels based on P2b and on neurofilaments

Characteristic gel property	P2b	Neurofilaments[3]
Bundle diameter, d_B	7.5 nm*	10 nm
Average bundle number, N	9	4
Persistence length†, $l_{p,B}$	460 nm	600 nm
Deformation regime ($G_0 \propto c^n$)	Entropic ($G_0 \propto c^{2.2}$)	Entropic ($G_0 \propto c^{2.5}$)
G_0†	100–1,000 Pa‡	2–20 Pa§
High-strain regime	Strain stiffening ($K' \propto \sigma^{3/2}$)	Strain stiffening ($K' \propto \sigma^{3/2}$)
Contour length†, l_c	110 nm	300 nm

Properties given in the first column were determined at similar concentrations; exceptions are shown by footnotes.
* Calculated based on N and an estimated cross-section of the polymers.
† Determined at 1 mg ml^{-1}.
‡ Temperature range: 30 °C < T < 60 °C.
§ Mg^{2+} concentration range: 2 mM < [Mg^{2+}] < 20 mM.

intrinsic polymer properties related to the secondary structure of the chains. After determination of N, equations (1) and (2) provide the other unknown quantities: $l_{p,B} = 460$ nm and $l_c = 110$ nm (at 1 mg ml^{-1}, 30 °C); the latter is significantly smaller than $l_{p,B}$, as would be expected for a semi-flexible network. The gels of **P2b** closely resemble those formed by neurofilaments (a typical class of intermediate filaments), not only in their mechanical properties, but also in their characteristic length scales—for example, bundle diameter, pore size and bundle stiffness (Table 1).

The model has now been modified to write G_0 and σ_c at given experimental conditions as a function of the intrinsic (temperature dependent) single-chain persistence length, the bundle number and the length between crosslinks (see Supplementary Information):

$$G_0(T) \propto N^3 \frac{c}{l_c^3} RT l_{p,0}^2(T) \quad (3)$$

and

$$\sigma_c(T) \propto N \frac{c}{l_c^2} RT l_{p,0}(T) \quad (4)$$

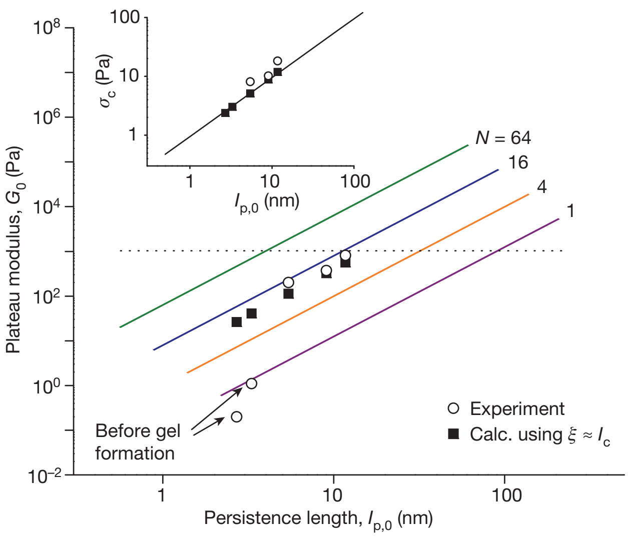


Figure 4 | Stiffness of the gel versus stiffness of the constituent polymer. Main panel: open circles show G_0 of **P2b** as a function of $l_{p,0}(T)$ at $T = 10$–60 °C and $c = 1$ mg ml^{-1}; filled squares show G_0 values of **P2b** calculated using equation (3) (substituting l_c for ξ) at the same temperatures and using $N = 9.1$. The coloured lines represent general trends, obtained from equation (3), which can be used to correlate $l_{p,0}$ to G_0 at a set concentration and given $N = 1$ – 64. The dotted line at $G_0 = 1$ kPa is shown for reference; it shows that 1-kPa gels can be prepared from a very stiff single polymer chain as well as from much more flexible, tightly bundled polymers (large N). The inset shows the variation of the calculated critical stress σ_c with $l_{p,0}$ (equation (4)), which is independent of N. The open circles are experimental data points obtained at $T = 30$, 40 and 50 °C. The corresponding calculated points (filled squares) overlap with the trend lines of $N = 1$ – 64.

Using equations (3) and (4) as a starting point, we can now speculate on how these hydrogels could be further engineered. For instance, is it possible to go even lower in concentration, can we set the pore size of a hydrogel, or can we generate stiffer gels that mimic the properties of the other cytoplasmic or extracellular materials?

To this end, we approximated the experimentally poorly accessible l_c (which scales with c as $l_c \propto c^{-0.4}$) by the mesh size ξ (which scales as $\xi \propto c^{-0.5}$, see Supplementary Information) that can be readily calculated from known molecular parameters, N and c ($\xi = 140$ nm for **P2b** at 1 mg ml^{-1}, 30 °C). When we further disregard the potential transition from the tight to the loose bundle regime, we can calculate G_0 as a function of the single chain persistence length $l_{p,0}$ and N (Fig. 4). The plot highlights that, even for intrinsically very stiff polymers, bundling is a prerequisite for good mechanical properties of the gel. Controlling bundling presents a central challenge for molecular chemists, because it allows tuning of both the gel modulus ($G_0 \propto \sqrt{N^3}$) and the pore size ($\xi \propto \sqrt{N}$) of the gel. This analysis is completely in line with how nature controls the mechanical properties of cytoskeletal soft materials: taking stiff protein elements (a variety of elements of different dimensions provide flexibility in the design) and controlling the amount of bundling by regulating the concentration of crosslinking proteins or divalent cations.

We have presented a truly artificial mimic of intermediate filaments, with all their characteristic mechanical properties. The helical polyisocyanide polymer plays a crucial role in providing an intrinsically stiff backbone and controlling the bundling process. However, this class of materials goes beyond mimicking intermediate-filament biogels, because network characteristics can be readily manipulated through small modifications in the chemical structure—for instance, gel transition temperatures can be changed by the length of the ethylene glycol tail and the intrinsic backbone stiffness by the amino acid sequence[6,27]. Moreover, functional groups can be introduced at the periphery of the polymer which allows for the incorporation of a wide variety of (bio-)molecules or cross-linkers in the polymer, mimicking more closely the natural environment of the cell.

METHODS SUMMARY

Materials. The polymerization of **1–3** was carried out with $Ni(ClO_4)_2 \cdot 2H_2O$ as catalyst in toluene. The reaction mixtures were stirred vigorously in a sealed flask at room temperature for two hours. The solvent was removed and the residue was precipitated three times from chloroform or tetrahydrofuran in diethyl ether. The products were routinely characterized with infrared and circular dichroism spectroscopies, and AFM. NMR spectroscopy gives broad signals only.

AFM analysis. To visualize individual polymer chains, solutions (~1 μM in $CHCl_3$) were spin-coated on freshly cleaved muscovite mica substrates. Polymer gels were deposited on the substrate by direct contact transfer. All images were obtained by tapping mode AFM.

SMFS. Before analysis, the AFM tip was cleaned meticulously. Polymer samples (3 mM in CH_2Cl_2) were spin-coated on freshly cleaved muscovite mica. The substrate was rinsed with MilliQ water to remove non-absorbed polymer. After morphology and density characterization (tapping mode AFM in air), the SMFS measurements were conducted in MilliQ water. Owing to the low density on the surface, less than 1% of approach–retract cycles yielded successful traces.

Rheology. Samples were dissolved with regular vortexing in (cold) demineralized water at least 24 h before the measurements. Rheological measurements were

carried out in Couette geometry with heating/cooling rates of $2\,^{\circ}\mathrm{C\,min^{-1}}$. Standard measurements were carried out at 4% strain and at different frequencies (0.5–5 Hz). The data shown in the manuscript was recorded at 1 Hz. For each sample, this was in the linear response regime. Nonlinear rheology in the gel phase was carried out at 50 °C after equilibrating for 15 min using a pre-stress protocol[25]. A detailed description of all techniques and the modified semi-flexible network model is given in Supplementary Information.

Received 7 May; accepted 4 December 2012.
Published online 23 January 2013.

1. Kamm, R. D. & Mofrad, M. R. K. in *Cytoskeletal Mechanics: Models and Measurements* (eds Mofrad, M. R. K. & Kamm, R. D.) Ch. 1, 1–17 (Cambridge Univ. Press, 2006).
2. Fernández, P., Pullarkat, P. A. & Ott, A. A master relation defines the nonlinear viscoelasticity of single fibroblasts. *Biophys. J.* **90,** 3796–3805 (2006).
3. Fernandez-Gonzalez, R. & Zallen, J. A. Feeling the squeeze: live-cell extrusion limits cell density in epithelia. *Cell* **149,** 965–967 (2012).
4. Storm, C., Pastore, J. J., MacKintosh, F. C., Lubensky, T. C. & Janmey, P. A. Nonlinear elasticity in biological gels. *Nature* **435,** 191–194 (2005).
5. Schwartz, E., Le Gac, S., Cornelissen, J. J. L. M., Nolte, R. J. M. & Rowan, A. E. Macromolecular multi-chromophoric scaffolding. *Chem. Soc. Rev.* **39,** 1576–1599 (2010).
6. Cornelissen, J. J. L. M. *et al.* β-helical polymers from isocyanopeptides. *Science* **293,** 676–680 (2001).
7. Keereweer, B. *et al.* in *Functional Supramolecular Architectures* Vol. 1 (eds Samori, P. & Cacialli, F.) Ch. 5, 135–152 (VCH, 2011).
8. Lin, Y.-C. *et al.* Origins of elasticity in intermediate filament networks. *Phys. Rev. Lett.* **104,** 058101 (2010).
9. MacKintosh, F. C., Kas, J. & Janmey, P. A. Elasticity of semiflexible biopolymer networks. *Phys. Rev. Lett.* **75,** 4425–4428 (1995).
10. Gardel, M. L. *et al.* Elastic behavior of cross-linked and bundled actin networks. *Science* **304,** 1301–1305 (2004).
11. Tiller, J. C. Increasing the local concentration of drugs by hydrogel formation. *Angew. Chem. Int. Edn* **42,** 3072–3075 (2003).
12. Place, E. S., Evans, N. D. & Stevens, M. M. Complexity in biomaterials for tissue engineering. *Nature Mater.* **8,** 457–470 (2009).
13. Peppas, N. A., Hilt, J. Z., Khademhosseini, A. & Langer, R. Hydrogels in biology and medicine: from molecular principles to bionanotechnology. *Adv. Mater.* **18,** 1345–1360 (2006).
14. Hirst, A. R., Escuder, B., Miravet, J. F. & Smith, D. K. High-tech applications of self-assembling supramolecular nanostructured gel-phase materials: from regenerative medicine to electronic devices. *Angew. Chem. Int. Edn* **47,** 8002–8018 (2008).
15. Rowan, A. E. *et al.* Method for the preparation of high molecular weight oligo(alkylene glycol) functionalized polyisocyanopeptides. European Patent 2,287,221 (2011).
16. Grason, G. M. & Bruinsma, R. F. Chirality and equilibrium biopolymer bundles. *Phys. Rev. Lett.* **99,** 098101 (2007).
17. Pollard, T. D. & Cooper, J. A. Actin and actin-binding proteins — a critical evaluation of mechanisms and functions. *Annu. Rev. Biochem.* **55,** 987–1035 (1986).
18. Leterrier, J. F., Kas, J., Hartwig, J., Vegners, R. & Janmey, P. A. Mechanical effects of neurofilament cross-bridges — modulation by phosphorylation, lipids, and interactions with F-actin. *J. Biochem.* **271,** 15687–15694 (1996).
19. Han, S., Hagiwara, M. & Ishizone, T. Synthesis of thermally sensitive water-soluble polymethacrylates by living anionic polymerizations of oligo(ethylene glycol) methyl ether methacrylates. *Macromolecules* **36,** 8312–8319 (2003).
20. Lutz, J. F. & Hoth, A. Preparation of ideal PEG analogues with a tunable thermosensitivity by controlled radical copolymerization of 2-(2-methoxyethoxy)ethyl methacrylate and oligo(ethylene glycol) methacrylate. *Macromolecules* **39,** 893–896 (2006).
21. Wang, H. *et al.* A structure-gelation ability study in a short peptide-based 'Super Hydrogelator' system. *Soft Matter* **7,** 3897–3905 (2011).
22. Mason, T. G., Dhople, A. & Wirtz, D. Linear viscoelastic moduli of concentrated DNA solutions. *Macromolecules* **31,** 3600–3603 (1998).
23. Onck, P. R., Koeman, T., van Dillen, T. & van der Giessen, E. Alternative explanation of stiffening in cross-linked semiflexible networks. *Phys. Rev. Lett.* **95,** 178102 (2005).
24. Huisman, E. M., van Dillen, T., Onck, P. R. & Van der Giessen, E. Three-dimensional cross-linked F-actin networks: relation between network architecture and mechanical behavior. *Phys. Rev. Lett.* **99,** 208103 (2007).
25. Broedersz, C. P. & MacKintosh, F. C. Molecular motors stiffen non-affine semiflexible polymer networks. *Soft Matter* **7,** 3186–3191 (2011).
26. Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of λ-phage DNA. *Science* **265,** 1599–1600 (1994).
27. Van Buul, A. M. *et al.* Stiffness versus architecture of single helical polyisocyanopeptides. *Chem. Sci*. (submitted).
28. Bathe, M., Heussinger, C., Claaessens, M. M. A. E., Bausch, A. R. & Frey, E. Cytoskeletal bundle mechanics. *Biophys. J.* **94,** 2955–2964 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Norder for assistance with rheological experiments, C. Broersz for support with nonlinear rheology, F. MacKintosh for discussions on the interpretation of the semi-flexible polymer network theory and E. Cator for work on the statistical analysis of the AFM images. We acknowledge financial support from Technology Foundation STW, the Council for the Chemical Sciences of the Netherlands Organisation for Scientific Research (NWO-CW-7005644), NRSCC, the Royal Academy for Arts and Sciences and EU projects Hierarchy (PITN-CT-2007-215851) and Superior (PITN-CT-2009-238177).

Author Contributions P.H.J.K. and A.E.R. wrote the manuscript and developed the model. M.K, Z.H.E.-A., T.W., E.S., H.J.K. and R.H. were involved in the design, synthesis and characterization of the materials. M.J. and A.M.v.B. conducted the SMFS measurements. V.A.A.L.S., P.H.J.K., E.M. and S.J.P. designed, conducted and interpreted the rheological experiment. P.H.J.K., R.J.M.N. and A.E.R. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.H.J.K. (p.kouwer@science.ru.nl) or A.E.R. (a.rowan@science.ru.nl).

doi:10.1038/nature11784

Divergent global precipitation changes induced by natural versus anthropogenic forcing

Jian Liu[1,2], Bin Wang[3], Mark A. Cane[4], So-Young Yim[3] & June-Yi Lee[3]

As a result of global warming, precipitation is likely to increase in high latitudes and the tropics and to decrease in already dry subtropical regions[1]. The absolute magnitude and regional details of such changes, however, remain intensely debated[2,3]. As is well known from El Niño studies, sea-surface-temperature gradients across the tropical Pacific Ocean can strongly influence global rainfall[4,5]. Palaeoproxy evidence indicates that the difference between the warm west Pacific and the colder east Pacific increased in past periods when the Earth warmed as a result of increased solar radiation[6–9]. In contrast, in most model projections of future greenhouse warming this gradient weakens[2,10,11]. It has not been clear how to reconcile these two findings. Here we show in climate model simulations that the tropical Pacific sea-surface-temperature gradient increases when the warming is due to increased solar radiation and decreases when it is due to increased greenhouse-gas forcing. For the same global surface temperature increase the latter pattern produces less rainfall, notably over tropical land, which explains why in the model the late twentieth century is warmer than in the Medieval Warm Period (around AD 1000–1250) but precipitation is less. This difference is consistent with the global tropospheric energy budget[12], which requires a balance between the latent heat released in precipitation and radiative cooling. The tropospheric cooling is less for increased greenhouse gases, which add radiative absorbers to the troposphere, than for increased solar heating, which is concentrated at the Earth's surface. Thus warming due to increased greenhouse gases produces a climate signature different from that of warming due to solar radiation changes.

How much will precipitation increase as the world warms as a result of increased greenhouse gases[2,3,12]? Will the greenhouse-warming-induced precipitation change be different from that induced by natural forcing? Past climate changes might provide guidance. Much has been achieved in the reconstruction of climate from proxy data (tree ring, stalagmites, ice cores, corals, laminated sediments and historical documents) and in numerical simulations of climate change over the past thousand years[13]. There has been great progress in understanding millennial variations of global mean temperature[14,15] and dynamical modes of climate variability such as the North Atlantic Oscillation and the El Niño–Southern Oscillation (ENSO)[7], but knowledge of precipitation change remains quite limited and primarily confined to regional scales[16,17].

Here we examine differences over the last millennium between global precipitation changes due to natural changes in the solar–volcanic forcing, that is, the sum of the radiative effects of variations in solar irradiance and volcanic aerosols, and precipitation changes resulting from greenhouse-gas forcing. Because proxy data are sparse and the spatial distribution of precipitation is complex, our approach relies on millennial simulations with ECHO-G, an atmosphere–ocean coupled climate model able to reproduce realistic present-day climatology and short-term climate fluctuations (Supplementary Information). The model-simulated present-day precipitation climatology is comparable to those derived from the state-of-the-art reanalysis data (Supplementary Fig. 1) or the climate models with the best precipitation simulations (see Supplementary Fig. 2). Investigations with this model of various aspects of climate variability including temperature, ENSO and global monsoons[18–20] have built confidence in the model's credibility for understanding physical processes pertinent to global precipitation change. It is important to note that the simulation we study treats the effect of volcanic aerosol as if it were exactly the same as a reduction in solar radiance[21]. Because it is difficult to extract an unambiguous pattern for the response to greenhouse gases from a simulation that ends in the twentieth century and that also includes solar forcing, we turn to results from a simulation with the same ECHO-G model of the twenty-first century forced by the A1B scenario of greenhouse-gas increases and from a simulation forced by observed greenhouse gases only from 1860–2000 (Supplementary Information).

Strikingly, although the late twentieth century is warmer than the Medieval Warm Period, rainfall is less (Fig. 1). How much global

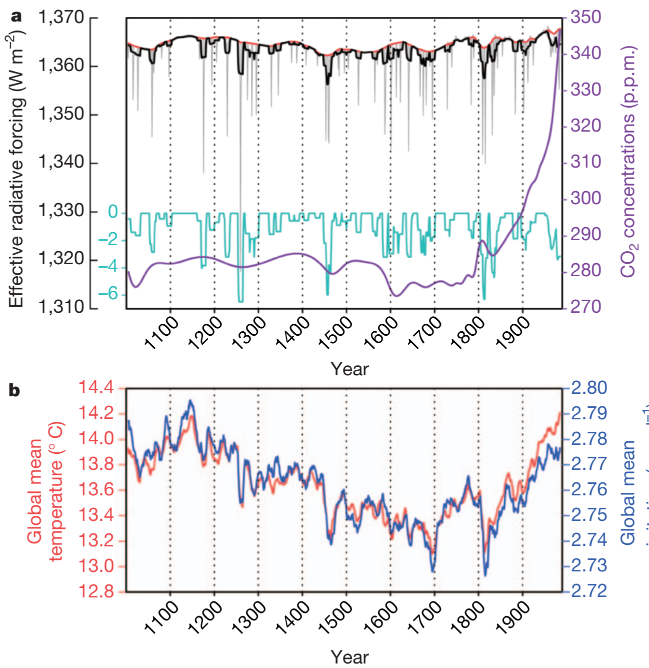


Figure 1 | The external forcing and responses. a, The grey line shows the annual mean time series of effective radiative (solar and volcanic) forcing. The red line shows the 11-year running mean time series of solar radiation. The blue line shows volcanic radiative forcing. The black line shows the effective radiative (solar-volcanic) forcing. The purple line shows the CO_2 concentration (right axis). **b**, Shown are the global mean temperature (red), and the global mean precipitation intensity (blue) simulated in the forced run with the ECHO-G model. (p.p.m., parts per million.)

[1]Key Laboratory of Virtual Geographic Environment of Ministry of Education, School of Geography Science, Nanjing Normal University, Nanjing 210023, China. [2]State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China. [3]International Pacific Research Center and Department of Meteorology, University of Hawaii at Manoa, Honolulu, Hawaii 96825, USA. [4]Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA.

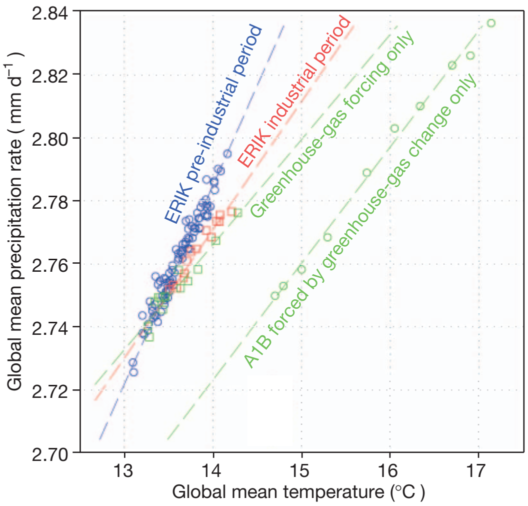


Figure 2 | Scatter plot of decadal means of the global mean precipitation rate versus the global mean temperature, at 2 m above the surface. Blue circles show the millennium simulation (ERIK) for the pre-industrial period (AD 1000–1850), with regression slope 2.1% °C^{-1} (0.058 mm d^{-1} °C^{-1}). Red squares show ERIK for the industrial period (1850–1990), with regression slope 1.4% °C^{-1} (0.039 mm d^{-1} °C^{-1}). Green symbols show two ECHO-G model runs with only greenhouse-gas forcing, with regression slopes 1.2% °C^{-1} (0.033 mm d^{-1} °C^{-1}) (squares) and 1.3% °C^{-1} (0.036 mm d^{-1} °C^{-1}) (circles).

precipitation would increase for a given temperature increase due to global warming has been the subject of intense debate[2,3]. In the forced millennium simulation (Supplementary Information) this ratio is about 2.1% °C^{-1} during the pre-industrial period (AD 1000–1850), but only about 1.4% °C^{-1} during the industrial period (AD 1850–1990) (Fig. 2), a difference that is significant above the 95% confidence level (Supplementary Information). Figure 2 also shows that in two runs with the same model forced only by greenhouse gases (Methods), the ratio is close to but less than that for the industrial period and is again distinct from that in the pre-industrial period when the only significant forcing is solar. Regional precipitation changes depend on circulation changes and are influenced by local sea surface temperature (SST), and the global precipitation change is not solely dependent on global mean temperature, but what accounts for the difference in this ratio between the Medieval Warm Period and the present?

A good starting point is the tropospheric energy budget: whereas the change in atmospheric water vapour is closely controlled by temperature via the Clausius–Clapeyron relation, precipitation changes are constrained by the energy budget[12]. For the troposphere as a whole, the precipitation heating is principally balanced by radiative flux divergence. Thus the total global precipitation is controlled by the difference between the upward radiative flux at the tropopause and that at the earth's surface. All other things being equal, an increase in surface temperature, whether it is due to increased solar flux or increased greenhouse trapping, will increase this flux divergence and hence increase precipitation. However, adding long-wave absorbers to the atmosphere will tend to lessen the difference between the flux at the top and that at the bottom, so the increase in precipitation will be less than if the surface heating results from increased solar radiation[12].

The energy argument can explain the difference in global mean precipitation, but it does not address the spatial distribution of precipitation changes. We first estimate the changes in precipitation and SST by differencing these fields at a time of high solar radiance and little volcanic aerosol (AD 1100–1200, during the Medieval Warm Period; see Fig. 1a) and a time of low solar radiance and high volcanic aerosol (AD 1630–1730, during the Little Ice Age). We take 100-year averages to reduce the influence of higher-frequency natural variability. We will refer to the derived mode as the solar–volcanic mode. It features an enhanced zonal SST gradient in the tropical Pacific Ocean (Fig. 3a) and, as might be expected, the stronger SST gradient is accompanied by stronger easterlies in the equatorial Pacific, and a stronger Walker circulation. It has these features in common with a La Niña event, but the pattern differs in many respects, including having a positive value of the ENSO index NINO3.4 (the SST anomaly averaged over the eastern equatorial box 5° S–5° N, 120° W–170° W; Supplementary Fig. 3), which would be negative for a La Niña event.

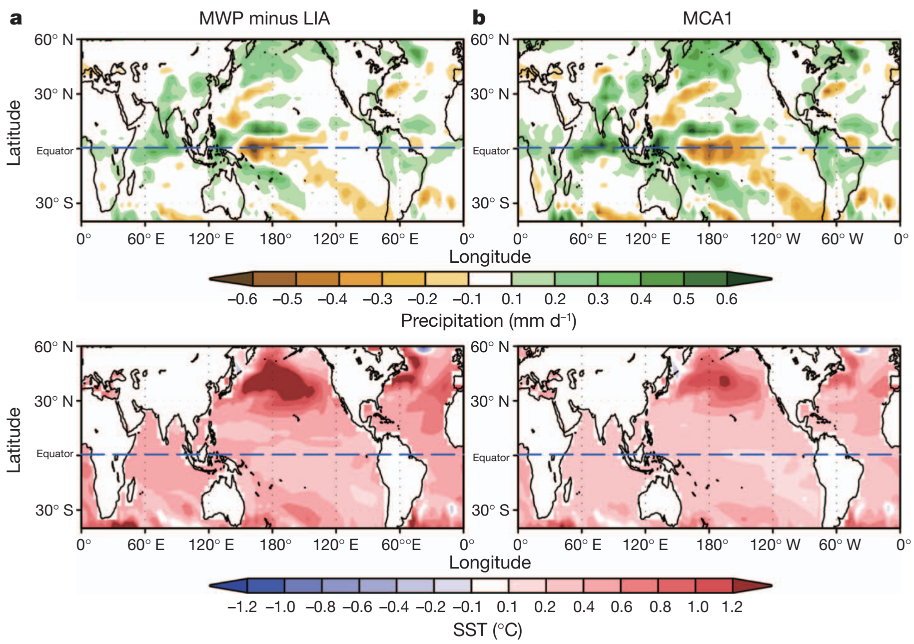


Figure 3 | Spatial patterns of the solar–volcanic forced mode. **a**, The precipitation and SST changes are shown for the Medieval Warm Period (MWP, AD 1100–1200) minus the Little Ice Age (LIA, AD 1630–1730) in response to differences in solar forcing. **b**, The precipitation and SST patterns of the leading maximum covariance analysis mode in the ERIK millennium run are shown for the period AD 1000–1850, based on 11-year running means after the leading mode of internal variability is removed. (They explain 15.3% and 11.1% of the variance, respectively.) The pattern correlation coefficients between **a** and **b** over the entire domain are 0.92 for precipitation and 0.98 for SST.

We checked this result with a different technique for detecting major patterns of decadal precipitation variations. To focus on the forced response, we first removed the leading internal mode component from the millennium run (Methods and Supplementary Figs 3 and 4) and then applied a maximum covariance analysis (that is, a singular value decomposition[22]) to the precipitation and SST fields for the period AD 1000–1850, when the only appreciable forcing is solar–volcanic. The leading coupled spatial patterns of SST and precipitation (Fig. 3b) are markedly similar to the strong gradient patterns of the solar–volcanic mode (Fig. 3a); spatial correlation coefficients are 0.98 and 0.92 for the SST and precipitation fields, respectively. The time expansion coefficients of the precipitation and SST (Supplementary Fig. 5) show a prominent centennial–millennial fluctuation, with a substantial dry and cold epoch during the Little Ice Age (AD 1450–1850), when radiance at the surface due to solar–volcanic forcing is low, and a wet and warm epoch occurring in the Medieval Warm Period (AD 1000–1250), when radiance is high.

The implication that the Medieval Warm Period featured a solar–volcanic pattern (in particular, a stronger zonal SST gradient) but that the global cooling at the Little Ice Age has the opposite pattern agrees with available proxy evidence[6,7,9] and model results[23–25]. With increased solar–volcanic forcing the rainfall increases over the climatological 'wet' regions, resulting in an overall increase in global mean precipitation (Fig. 3).

To estimate the response induced by greenhouse-gas forcing, we examined two greenhouse-gas-only forcing runs, one for the industrial period (AD 1860–2000) with observed greenhouse-gas concentration as the only forcing, and the other for AD 1990–2100, forced by the A1B scenario of increased greenhouse gases (Supplementary Fig. 6). The resultant trend patterns of SST and precipitation for the two runs are similar except that the A1B run has substantially larger amplitudes than the industrial run, owing to stronger greenhouse-gas forcing. Figure 4a shows the greenhouse-gas mode estimated from the ECHO-G A1B run, which is similar to the Coupled Model Intercomparison Project Phase 5 (CMIP5) multi-model mean projection (Fig. 4b), showing that this pattern in response to greenhouse-gas forcing is common among models. In contrast to the strong zonal SST gradient forced by solar warming, this greenhouse-gas forced mode shows a reduced equatorial Pacific zonal SST gradient. Corresponding to the enhanced and reduced zonal SST gradients, the overall increase of global mean precipitation in the solar–volcanic mode is larger than that in the greenhouse-gas mode, though the solar–volcanic mode is drier than the greenhouse-gas mode in the central equatorial Pacific. The global total precipitation increase for a given temperature increase due to greenhouse-gas warming (about 1.2% to 1.3% per °C) is about 40% less than that due to solar–volcanic warming (2.1% per °C) (Fig. 2). Of note for societal impacts is that for the solar–volcanic forced mode the rainfall over tropical land increases by 5.5% for a 1 °C increase in global mean temperature, while for the greenhouse-gas forced mode the corresponding increase of 2.4% is less than half of that.

The late twentieth century is warmer than the Medieval Warm Period but the rainfall is less (Fig. 1b), and we argued above that the difference may be attributed to the difference in energy budget constraints when the warming is due to increased greenhouse gases as opposed to solar–volcanic heating. The climate system accommodates the energy budget differences by changing the pattern of global warming. For the same increase in global mean temperature, the solar forced pattern has a stronger SST gradient than the greenhouse-gas forced pattern. Along with the enhanced SST gradient, the Walker circulation strengthens and moisture convergence is concentrated in the Indo-Pacific warm pool region. This wet region gets wetter[1,2], augmenting global precipitation.

There is ongoing debate about whether the equatorial Pacific responds to increased heating by enhancing the east–west gradient[23,24] or reducing it[2,10,11]. The "ocean dynamical thermostat" theory[23,24,26] argues that increased heating at the surface warms SSTs in the west more because in the east the heating is countered by upwelling of cold waters from below. The increase in SST gradient gives rise to an enhanced pressure gradient and hence stronger easterly winds and a stronger Walker circulation, which in turn enhance the SST gradient, a mechanism known as "the Bjerknes feedback". The essence of the opposing argument[2,11] is that because the warming increases the moist static energy in the atmosphere by a greater amount than the energy transports associated with precipitation, the Walker circulation must slow down, so the Bjerknes feedback now implies a weaker SST gradient. The "ocean dynamical thermostat" argument draws support from palaeoclimate proxy data[6–9,27] and intermediate model simulations[23,24]. On the other hand, Intergovernmental Panel on Climate Change (IPCC) model projections for the twenty-first century typically show a weaker zonal SST gradient[1,25], supporting the weaker Walker argument.

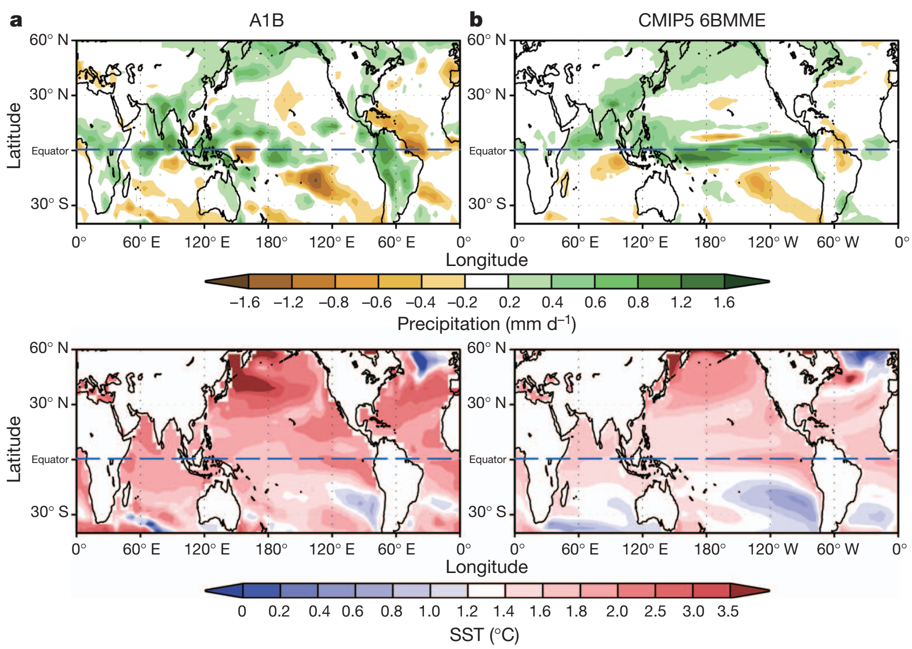


Figure 4 | Comparison of the changes in annual mean precipitation and SST. **a**, The ECHO-G simulated changes under the A1B run forced by greenhouse-gas change only for the period AD 2070–2099 relative to the period AD 1990–2019. **b**, The change in precipitation and SST for the period AD 2070–2099 relative to AD 1980–2005 in the multi-model mean of the six best (6BMME) CMIP5 models forced according to Representative Concentration Pathway 4.5 (a scenario that stabilizes radiative forcing at 4.5 W m^{-2} or less in the year AD 2100) forcing (ref. 30). The six best models are the ones with the best simulation of the mean and annual cycle precipitation, as shown in Supplementary Fig. 2. The pattern correlations between **a** and **b** are 0.42 for precipitation and 0.98 for SST.

Figures 3 and 4 show that solar heating led to a stronger SST gradient, whereas greenhouse-gas heating led to a weaker one, demonstrating that there is no contradiction between the palaeoclimate records and the IPCC simulations and that both theories may have a realm of validity. However, neither theory indicates that the outcome should depend on the type of heating. We found that, consistent with the earlier argument for less precipitation with greenhouse-gas forcing than with solar–volcanic forcing, the increase in atmospheric static stability is noticeably greater with greenhouse-gas forcing (Supplementary Fig. 7). The increased atmospheric stability favours a weaker zonal circulation and the accompanying weaker SST gradient characterizing the greenhouse-gas mode. We suggest that although the thermostat and associated stronger gradient pattern dominated in the past when the external warming was solar–volcanic, the weaker gradient pattern associated with greenhouse-gas forcing will dominate future change.

METHODS SUMMARY

Two millennial simulations[28] and two greenhouse-gas-only forcing runs with the ECHO-G coupled climate model[29] were analysed: (1) a 1,000-year control (free) simulation generated using fixed annually cycling forcing set at present-day values; (2) a forced run, named ERIK, covering the period AD 1000–1990, which is externally forced by solar variability, the effective radiative effects from stratospheric volcanic aerosols, and greenhouse-gas concentrations in the atmosphere, including CO_2 and CH_4, for the period AD 1000–1990; (3) a greenhouse-gas forced run for the period AD 1860–2000 with initial conditions selected from a long pre-industrial control simulation. Nineteen observed greenhouse gases were used, including CO_2, CH_4 and N_2O (ref. 18); and (4) an emissions scenarios (SRES) balance across all sources (A1B) run from AD 1990 to 2100 with 720 p.p.m. stabilization at AD 2100.

To identify the internal decadal variation mode, a principal-component analysis of the 11-year running mean precipitation was performed. To detect major patterns of forced decadal variation, we first removed the leading internal mode component of the precipitation from the ERIK run, and then applied a maximum covariance analysis[22] to the precipitation and SST fields for the period AD 1000–1850. To test whether the difference between the two slopes (Fig. 2) was due to sampling errors, we used the Student's *t*-test (it was not). Further details are given in the Supplementary Information.

Received 21 September; accepted 8 November 2012.

1. Meehl, G. A. *et al.* Global climate projections. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 747–845 (Cambridge Univ. Press, 2007).
2. Held, I. M. & Soden, B. J. Robust responses of the hydrological cycle to global warming. *J. Clim.* **19,** 5686–5699 (2006).
3. Wentz, F., Ricciardulli, L., Hilburn, K. & Mears, C. How much more rain will global warming bring? *Science* **317,** 233–235 (2007).
4. Ropelewski, C. F. & Halpert, M. S. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Weath. Rev.* **115,** 1606–1626 (1987).
5. Ropelewski, C. F. & Halpert, M. S. Quantifying Southern Oscillation–precipitation relationships. *J. Clim.* **9,** 1043–1059 (1996).
6. Adams, J. B., Mann, M. E. & Ammann, C. M. Proxy evidence for an El Niño-like response to volcanic forcing. *Nature* **426,** 274–278 (2003).
7. Cobb, K. M., Charles, C. D., Cheng, H. & Edwards, R. L. El Niño-Southern Oscillation and tropical Pacific climate during the last millennium. *Nature* **424,** 271–276 (2003).
8. Mann, M. E., Cane, M. A., Zebiak, S. E. & Clement, A. Volcanic and solar forcing of the tropical Pacific over the past 1000 years. *J. Clim.* **18,** 447–456 (2005).
9. Mann, M. E. *et al.* Global signatures and dynamical origins of the little ice age and medieval climate anomaly. *Science* **326,** 1256–1260 (2009).
10. Meehl, G. A. & Washington, W. M. El Niño like climate change in a model with increased atmospheric CO_2 concentration. *Nature* **382,** 56–60 (1996).
11. Vecchi, G. A. *et al.* Weakening of tropical Pacific atmospheric circulation due to anthropogenic forcing. *Nature* **441,** 73–76 (2006).
12. Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrologic cycle. *Nature* **419,** 224–232 (2002).
13. Mann, M. E. Climate over the past two millennia. *Annu. Rev. Earth Planet. Sci.* **35,** 111–136 (2007).
14. Mann, M. E., Bradley, R. S. & Hughes, M. K. Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophys. Res. Lett.* **26,** 759–762 (1999).
15. Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M. & Karlen, W. Highly variable northern hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature* **433,** 613–617 (2005).
16. Pauling, A., Luterbacher, J., Casty, C. & Wanner, H. 500 years of gridded high resolution precipitation reconstructions over Europe and the connection to large scale circulation. *Clim. Dyn.* **26,** 387–405 (2006).
17. Dore, M. H. I. Climate change and changes in global precipitation patterns: what do we know? *Environ. Int.* **31,** 1167–1181 (2005).
18. Min, S. K. & Hense, A. A Bayesian assessment of climate change using multimodel ensembles. Part I: global mean surface temperature. *J. Clim.* **19,** 3237–3256 (2006).
19. Rodgers, K. B., Friedriches, P. & Latif, M. Tropical Pacific decadal variability and its relationship to decadal modulations of ENSO. *J. Clim.* **17,** 3761–3774 (2004).
20. Liu, J. *et al.* Centennial variations of the global monsoon precipitation in the last millennium: results from the ECHO-G model. *J. Clim.* **22,** 2356–2371 (2009).
21. von Storch, H. *et al.* Reconstructing past climate from noisy data. *Science* **306,** 679–682 (2004).
22. Wallace, J. M., Smith, C. & Bretherton, C. S. Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J. Clim.* **5,** 561–576 (1992).
23. Clement, A. C., Seager, R., Cane, M. A. & Zebiak, S. E. An ocean dynamical thermostat. *J. Clim.* **9,** 2190–2196 (1996).
24. Cane, M. A. *et al.* 20th century sea surface temperature trends. *Science* **275,** 957–960 (1997).
25. Vecchi, G. A., Clement, A. & Soden, B. J. Examining the tropical Pacific's response to global warming. *Eos* **89,** 81 (2008).
26. Bauer, E., Claussen, M. & Brovkin, V. Assessing climate forcings of the Earth system for the past millennium. *Geophys. Res. Lett.* **30,** 1276 (2003).
27. Emile-Geay, J., Seager, R., Cane, M. A., Cook, E. R. & Haug, G. H. Volcanoes and ENSO over the last millennium. *J. Clim.* **21,** 3134–3148 (2008).
28. Zorita, E., Gonzalez-Rouco, J. F., von Storch, H., Montavez, P. & Valero, F. Natural and anthropogenic modes of surface temperature variations in the last thousand years. *Geophys. Res. Lett.* **32,** L08707 (2005).
29. Legutke, S. & Voss, R. *The Hamburg Atmosphere-Ocean Coupled Circulation Model ECHO-G.* Technical Report 18, 1–62 (German Climate Computer Center (DKRZ), 1999).
30. Lee, J. Y. & Wang, B. Future change of global monsoon in the CMIP5. *Clim. Dyn.* doi:10.1007/s00382-012-1564-0 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Basic Research Program (award numbers 2010CB950102 and XDA05080800 to J.L.) and Natural Science Foundation of China (award number 40871007 to J.L. and B.W.). B.W. and J.-Y.L. acknowledge the Global Research Laboratory (GRL) Program from the Korean Ministry of Education, Science and Technology (MEST, 2011-0021927). M.A.C. was supported by grant DE-SC0005108 from the Department of Energy and NOAA grant NA08OAR4320912. B.W., S.-Y.Y. and J.-Y.L. acknowledge support from the International Pacific Research Center, which is funded jointly by JAMSTEC, NOAA and NASA. We thank E. Zorita for providing ECHO-G millennium run data, and A. Hense and S.-K. Min for providing ECHO-G A1B and greenhouse-gas run data.

Author Contributions J.L. initiated the research. J.L., B.W. and M.A.C. contributed to the research and wrote the manuscript. S.-Y.Y. and J.-Y.L. made analyses and contributed to the graphics.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.C. (mcane@ldeo.columbia.edu) or J.L. (jliu@njnu.edu.cn).

LETTER

doi:10.1038/nature11818

Deep instability of deforested tropical peatlands revealed by fluvial organic carbon fluxes

Sam Moore[1]†, Chris D. Evans[2], Susan E. Page[3], Mark H. Garnett[4], Tim G. Jones[5], Chris Freeman[5], Aljosja Hooijer[6], Andrew J. Wiltshire[7], Suwido H. Limin[8] & Vincent Gauci[1]

Tropical peatlands contain one of the largest pools of terrestrial organic carbon, amounting to about 89,000 teragrams[1] (1 Tg is a billion kilograms). Approximately 65 per cent of this carbon store is in Indonesia, where extensive anthropogenic degradation in the form of deforestation, drainage and fire are converting it into a globally significant source of atmospheric carbon dioxide[1–3]. Here we quantify the annual export of fluvial organic carbon from both intact peat swamp forest and peat swamp forest subject to past anthropogenic disturbance. We find that the total fluvial organic carbon flux from disturbed peat swamp forest is about 50 per cent larger than that from intact peat swamp forest. By carbon-14 dating of dissolved organic carbon (which makes up over 91 per cent of total organic carbon), we find that leaching of dissolved organic carbon from intact peat swamp forest is derived mainly from recent primary production (plant growth). In contrast, dissolved organic carbon from disturbed peat swamp forest consists mostly of much older (centuries to millennia) carbon from deep within the peat column. When we include the fluvial carbon loss term, which is often ignored, in the peatland carbon budget, we find that it increases the estimate of total carbon lost from the disturbed peatlands in our study by 22 per cent. We further estimate that since 1990 peatland disturbance has resulted in a 32 per cent increase in fluvial organic carbon flux from southeast Asia—an increase that is more than half of the entire annual fluvial organic carbon flux from all European peatlands. Our findings emphasize the need to quantify fluvial carbon losses in order to improve estimates of the impact of deforestation and drainage on tropical peatland carbon balances.

Peatlands have high water tables and consequent low decomposition rates, and hence form large carbon stores[4]. Southeast Asian peat swamp forests (PSFs) currently experience extensive anthropogenic degradation in the form of deforestation, drainage and associated fire, all of which convert carbon stored in peat into atmospheric carbon dioxide (CO_2) via either direct combustion or through oxidation within the peat column[2,3]. Unlike boreal and temperate forests[5,6] and higher-latitude wetlands[7], however, the loss of fluvial organic carbon from tropical peats has yet to be fully quantified.

To quantify the effect of peatland degradation on fluvial organic C loss, we monitored dissolved organic carbon (DOC) and particulate organic carbon (POC) concentrations and water discharge rates from channels draining areas of both intact and disturbed PSF in a portion of central Kalimantan (Indonesia, Borneo) affected by severe deforestation, drainage and fire associated with the implementation of the Mega Rice Project. Initiated in 1995, this was a failed agricultural development project which aimed to convert one million hectares of peatland into rice fields[8]. We selected three PSF land-cover classes that differed in their recent disturbance history, located in or near to the Sebangau River basin (Supplementary Fig. 1 and Supplementary Information): (1) intact PSF (PSF1) (three channels in the Sebangau forest), (2) moderately drained disturbed PSF (PSF2) (two channels in Tubangnusa) and (3) severely drained disturbed PSF (PSF3) (three channels in Kalampangan). All disturbed PSF catchments to the east of the Sebangau River were comprised of lowland PSF of similar topography, peat thickness and vegetation to PSF1 before the Mega Rice Project disturbance[9], and experienced similar annual rainfall (Table 1).

Total organic carbon (TOC = DOC + POC) fluxes were monitored from each channel outlet at weekly intervals from June 2008 to May 2009. Results demonstrate larger mean annual TOC fluxes in both PSF2 and PSF3 (105 $g\,C\,m^{-2}\,yr^{-1}$ and 88 $g\,C\,m^{-2}\,yr^{-1}$, respectively) than in PSF1 (63 $g\,C\,m^{-2}\,yr^{-1}$; Fig. 1). This represents a 55% increase in TOC export from the disturbed sites (PSF2 and PSF3) over PSF1. Of the annual TOC flux from each land-cover class, 94% was lost during the wet season (October–June), the result of higher measured discharge rates (3.9 $m^3\,s^{-1}$ versus 1.0 $m^3\,s^{-1}$ in the dry season). This was associated with high rainfall rather than changes in carbon concentration, which remained relatively constant over the study period. As with seasonal variability, differences in discharge between land-cover classes dominated TOC flux variations, with higher discharge rates causing larger fluxes in PSF2 and PSF3 (1,744 mm and 1,724 mm, respectively) than in PSF1 (907 mm). These higher discharge rates in disturbed land-cover classes were not counterbalanced by lower TOC concentrations, and occurred despite uniform rainfall across sites (Table 1). This probably reflects a decline in evapotranspiration and increased runoff as a consequence of large-scale biomass loss and drainage in both disturbed land-cover classes (PSF2 and PSF3; see Methods). The DOC accounted for between 91–98% of the TOC lost, with lower DOC:POC ratios for disturbed sites (Table 1) suggesting that the drained and exposed peat is vulnerable to mechanical breakdown associated with the increased runoff.

Table 1 | Borneo study sites and land-cover class properties

Land-cover class	Number of channels	Area (km^2)	Rainfall (mm)	Total annual discharge (mm)	[Mean annual DOC] ($mg\,l^{-1}$)	[Mean annual POC] ($mg\,l^{-1}$)	[Mean annual TOC] ($mg\,l^{-1}$)	DOC:POC ratio	Annual TOC flux ($g\,C\,m^{-2}\,yr^{-1}$)
PSF1	3	34.2	2,810	907	68.0 ± 0.5	1.4 ± 0.04	69.5 ± 0.5	49:1	62.5 ± 0.70
PSF2	2	13.2	2,810	1,744	55.0 ± 1.0	5.3 ± 0.1	60.3 ± 0.9	10:1	105.3 ± 2.62
PSF3	3	64.0	2,810	1,724	48.3 ± 2.2	3.6 ± 0.1	51.9 ± 2.1	13:1	87.8 ± 3.88

Area means the total area of each land-cover class. Rainfall is the total annual rainfall, standardized using ground-station records and the "Tropical Rainfall Measuring Mission" satellite. The total annual discharge is standardized by area. DOC, POC and TOC concentrations and fluxes are shown as mean ± standard error of site means. Note that standard errors shown for TOC flux reflect concentration variations between sites only, because a single water yield was calculated per land-cover class. The DOC:POC ratio is [DOC]/[POC].

[1]Centre for Earth, Planetary, Space and Astronomical Research (CEPSAR), Department of Environment, Earth and Ecosystems, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK. [2]Centre for Ecology and Hydrology, Environment Centre Wales, Deiniol Road, Bangor LL57 2UW, UK. [3]Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, UK. [4]Natural Environment Research Council Radiocarbon Facility, Rankine Avenue, Scottish Enterprise Technology Park, East Kilbride G75 0QF, UK. [5]School of Biological Sciences, Bangor University, Deiniol Road, Gwynedd LL57 2UW, UK. [6]Deltares, PO Box 177, 2600 MH Delft, The Netherlands. [7]Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. [8]CIMTROP, University of Palangka Raya, Palangka Raya, Central Kalimantan, 73112, Indonesia. †Present address: Environment Change Institute, School of Geography and the Environment, University of Oxford, OX1 3QY Oxford, UK.

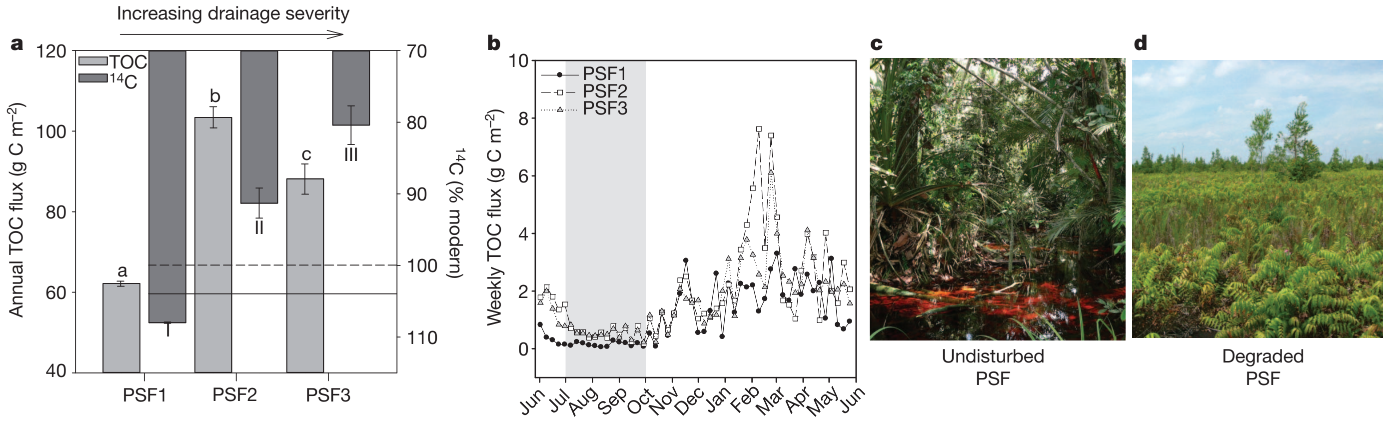


Figure 1 | Total and seasonal fluvial organic carbon losses from intact (PSF1) and disturbed (PSF2 and PSF3) catchments. a, Annual TOC flux (±s.e.m.). 'a', 'b' and 'c' denote significant differences between land-cover classes ($P < 0.05$, unpaired, two-sample t-test) and mean radiocarbon (^{14}C) levels (±s.e.m.) measured in DOC (wet-season samples); 'I', 'II' and 'III' denote significant differences ($P < 0.05$, unpaired, two-sample t-test). The solid horizontal line (104% of the modern value) represents the current atmospheric $^{14}CO_2$ level; the dashed horizontal line (100% modern) represents the composition of the atmosphere in 1950, in the absence of any anthropogenic influences (that is, fossil fuel burning and above-ground nuclear testing). **b**, Weekly TOC flux from all land-cover classes from June 2008 to June 2009 (grey shading indicates the dry season). PSF1 TOC is the sum of fluxes from three channels, PSF2 is the sum of fluxes from two channels and PSF3 is the sum of fluxes from three channels, all divided by the total area of the land-cover class. **c**, Intact PSF. **d**, Disturbed PSF. Copyright for Fig. 1c, d, S.E.P.

Surface-water DOC can derive from multiple sources, ranging from recent photosynthates to decomposition or dissolution products from deep within the peat column. We used radiocarbon (^{14}C) measurements to evaluate whether increased fluvial carbon loss from disturbed sites was due to increased inputs of fresh material or the result of destabilization and loss from peat that had been accumulating since the Last Glacial Maximum[10]. Previous DO^{14}C measurements from waters draining intact, peat-dominated catchments in North America[11], Siberia[12] and Europe[13,14] commonly show enrichment of DOC with 'bomb' carbon (associated with above-ground nuclear testing in the 1950s and 1960s), suggesting that the bulk of DOC leached from these systems is of recent origin, probably dominated by carbon fixed from the atmosphere within the last ten years[15]. This implies that DOC export does not represent a major loss pathway for long-term stored carbon[14]. However, none of these studies specifically examined disturbed (for example, deforested and drained) peatlands, and to our knowledge no measurements of DO^{14}C from tropical peatlands, either pristine or disturbed, have previously been reported.

We collected samples for DO^{14}C analysis from all sites at which we determined TOC fluxes, in August 2008 (dry season) and May 2011 (wet season). Significant differences between classes of land cover were observed during both seasons (Fig. 1 and Table 2). DOC lost from intact PSF was ^{14}C-enriched in both seasons (averaging 110–108% modern; see Supplementary Information for a description of ^{14}C methods). If all carbon in these samples is assumed to be of post-bomb origin[15], this ^{14}C signature can be reproduced by a simple model with over 99% of DOC deriving from carbon fixed from the atmosphere within the last 50 years (Fig. 2c). In contrast, DOC from channels draining disturbed land-cover classes was ^{14}C-depleted, ranging from 98.9–75.5% of the modern value (equivalent to ^{14}C ages of 92–2,260 years before present, BP, where 'present' means 1950). This ^{14}C depletion was observed in both dry- and wet-season samples. These data indicate that the increased DOC fluxes from disturbed peatlands are derived from previously stable carbon stored within the peat column, and suggest that this loss of carbon from depth is occurring throughout the seasonal hydrologic cycle. Application of an age attribution model (Fig. 2d) suggests that two-thirds of DOC in runoff from the PSF3 site derives from peat carbon aged 500–5,000 years.

We also measured DO^{14}C from two channels draining oil palm plantations in peninsular Malaysia that were previously PSF. Approximately 28,000 km^2 of industrial plantations are found in peninsular Malaysia, Sumatra and Borneo[16], making them a major contributor to PSF deforestation in the region. These samples had even lower DO^{14}C levels of 59% and 67% of the modern value, corresponding to mean ages of 4,180 yr BP and 3,180 yr BP respectively (Table 2). To our knowledge, these are the oldest soil-derived natural surface-water DO^{14}C measurements reported. We do not have comparable TOC flux data from these sites, although we note that measured concentrations at the time of sampling were lower than in Borneo (Table 2), and that these had fallen markedly from our initial measurements at the site in 1995 (48.8 mg per litre for intact PSF formerly on the site and 64.3 mg per litre for recently planted oil palm).

Our findings demonstrate that destabilization of the peat column at depth is responsible for the increases in organic carbon fluxes we

Table 2 | Hydrological, DOC, DO^{14}C and qualitative data for Bornean and Malaysian study sites

Land-cover class	Number of channels	Season	Total rainfall (mm)	[Mean DOC] (mg l^{-1})	DO^{14}C (% modern)	DO^{14}C age (yr BP)	SUVA$_{254}$ (litres per mg C per m)	Aromaticity (%)
PSF1	3	Dry	172	62.0	109.1 ± 0.3	Modern	4.06 ± 0.04	30.10
	3	Wet	1,141	64.1	108.0 ± 0.1	Modern	4.03 ± 0.08	29.91
PSF2	2	Dry	263	62.4	97.7 ± 0.6	188 ± 47	3.95 ± 0.15	29.38
	2	Wet	1,018	54.7	91.3 ± 2.1	735 ± 179	3.69 ± 0.19	27.69
PSF3	3	Dry	100	39.1	85.0 ± 0.6	1,308 ± 54	4.00 ± 0.14	29.71
	3	Wet	922	47.9	80.4 ± 2.7	1,760 ± 268	3.94 ± 0.10	29.32
Malaysia (abandoned)	1	Dry	499	6.0	67.3	3,184	4.93	35.80
Malaysia (active)	1	Dry	499	13.3	59.4	4,183	4.14	30.60

Rainfall refers to the three months before sampling. All data are shown as the mean of all sampled sites (± standard error of site means where more than one site was sampled). The DO^{14}C age represents the mean (± standard error) of estimated ages for individual sites based on their %modern value (see Supplementary Fig. 2). Specific ultraviolet absorbance (SUVA$_{254}$) and percentage aromaticity data are means of ten samples collected at weekly intervals during the wet and dry season. SUVA$_{254}$ is an indicator of the relative aromaticity of aquatic humic substances and of DOC as a whole, with high aromaticity being indicative of a high degree of recalcitrance. 'Abandoned' and 'active' refers to oil palm plantations.

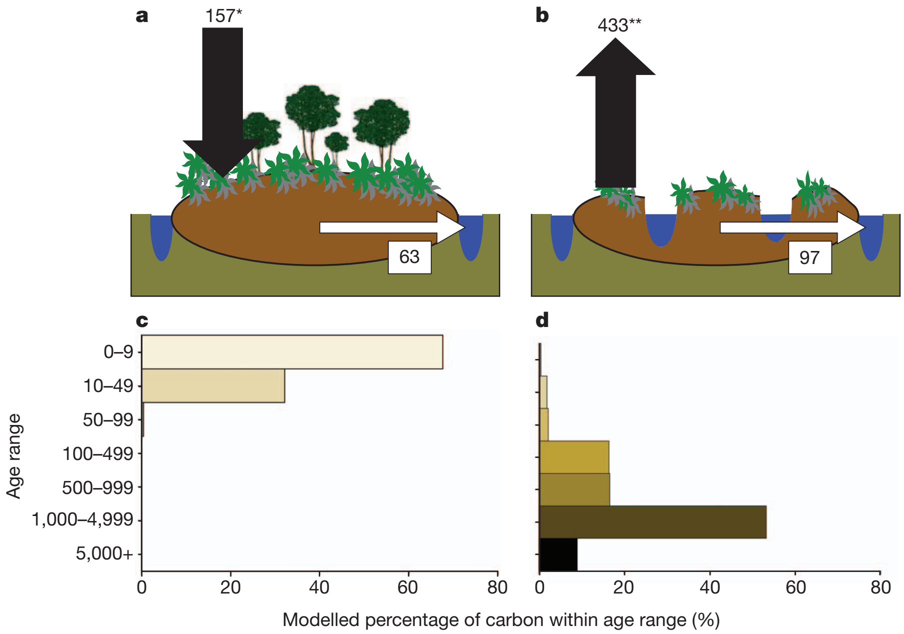


Figure 2 | Carbon balance and DOC age attribution of intact and disturbed PSF. **a**, **b**, Schematic showing net ecosystem exchange (black arrows; in grams of carbon in CO_2 per m^2 per year) and fluvial TOC loss (white arrows; in grams of carbon per m^2 per year) estimates in the PSF1 (**a**) and PSF2 and PSF3 (**b**) land-cover classes. *Net ecosystem exchange estimated from average 500-year estimate of carbon accumulation from a peat core taken within PSF1[17] to which mean fluvial carbon loss (63 $g\,C\,m^{-2}\,yr^{-1}$) has been added, thus approximating the net ecosystem exchange that would be measured by gaseous exchange alone. Carbon gain of intact PSF estimated to be 94 $g\,C\,m^{-2}\,yr^{-1}$ (net C sink). **Net ecosystem exchange measured from tower-based gaseous exchange measurements (eddy covariance) within the disturbed PSF catchments, which measures CO_2 fluxes across flat, deforested areas of the Mega Rice Project that are drained but do not contain drainage channels[18]. The net ecosystem carbon balance of disturbed PSF is estimated to be 530 $g\,C\,m^{-2}\,yr^{-1}$ (net C source). **c**, **d**, Modelled down-profile attribution of $DO^{14}C$ age from PSF1 (**c**) and PSF3 (**d**) land-cover classes respectively (wet season) as estimated from an age attribution model of $DO^{14}C$ age (see Supplementary Information for explanation).

observe from disturbed PSF. These large fluvial losses of old peat-derived carbon play an important part in altering the carbon balance of such ecosystems, yet because they are assumed to be small in comparison to gross primary productivity and ecosystem respiration, they are seldom measured. Although measurements of net ecosystem exchange for intact PSF are rare, we have a peat-core-derived carbon accumulation rate estimate of 94 $g\,C\,m^{-2}\,yr^{-1}$ from the PSF1[17] site (Fig. 2a). Including our fluvial C loss estimate of 63 $g\,C\,m^{-2}\,yr^{-1}$, this suggests an approximate gaseous net ecosystem exchange of -157 $g\,C\,m^{-2}\,yr^{-1}$ for PSF1. The measured net ecosystem exchange within our disturbed PSF sites is $+433$ $g\,C\,m^{-2}\,yr^{-1}$ (ref. 18), which results in an increased net ecosystem carbon balance of 530 $g\,C\,m^{-2}\,yr^{-1}$ (net carbon loss) if our intermediate disturbed PSF fluvial C loss estimate (97 $g\,C\,m^{-2}\,yr^{-1}$, the mean value for PSF2 and PSF3) is included. Thus, including fluvial C losses resulted in a 22% higher estimate of C loss from this disturbed site than was previously inferred from gaseous exchange measurements alone.

Applying our calculated TOC fluxes of 63 $g\,C\,m^{-2}\,yr^{-1}$ for intact PSF and our intermediate value of 97 $g\,C\,m^{-2}\,yr^{-1}$ for disturbed PSF, we estimate the annual TOC loss from the Sebangau basin (5,200 km^2) to be 0.41 $Tg\,C\,yr^{-1}$. This is within 10% of a basin-scale TOC loss estimate (0.46 $Tg\,C\,yr^{-1}$) for the River Sebangau[19] obtained during our study period. The broad agreement in flux estimates derived over contrasting scales gives us confidence that our calculated flux estimates for sub-catchments are representative of fluxes occurring at larger scales.

To quantify the impact peatland disturbance has had on regional long-term fluvial carbon loss, we applied our TOC flux estimates to land areas of intact and deforested PSF before and after peatland disturbance. We omitted industrial plantations from our calculations as, to our knowledge, there are no quantitative data on fluvial carbon flux from this land-cover class, although our $DO^{14}C$ data suggest that these ecosystems may also be highly unstable owing to land-use change. We estimate that since 1990, the conversion of intact PSF into disturbed peatland has resulted in around a 45% increase in the fluvial TOC flux, from 4.7 $Tg\,C\,yr^{-1}$ to 6.8 $Tg\,C\,yr^{-1}$ in Borneo, Sumatra and peninsular Malaysia, and a 32% (2.4 $Tg\,C\,yr^{-1}$) increase across the whole of southeast Asia (Table 3). This increase alone is more than half the entire annual European peatland fluvial organic carbon flux (4.3 $Tg\,C\,yr^{-1}$; estimated using a European peatland area of 292,000 km^2 (ref. 20) and an average fluvial carbon flux estimate of 14.6 $g\,C\,m^{-2}\,yr^{-1}$ (refs 21, 22)). Given the exclusion of peatland converted to plantations in our calculations, our estimated increase in regional fluvial organic carbon flux should be considered conservative.

The eventual fate of this additional fluvial carbon loss remains to be fully characterized, but it is known from other studies that most DOC is processed and emitted to the atmosphere as CO_2 and/or CH_4

Table 3 | Annual TOC fluxes pre- and post-disturbance at various spatial scales

Region	'Pre' dates	'Post' dates	Intact area 'pre' (km^2)	Intact area 'post' (km^2)	Disturbed area (km^2)	Total TOC flux 'pre' (Tg)	Total TOC flux 'post' (Tg)	Increase (Tg)
Mega Rice Project*	1991	2000	15,604	11,102	4,502	1.0	1.2	0.2
Borneo, Sumatra and Peninsular Malaysia	1990	2008	75,805†	15,600	60,205	4.7	6.8	2.1
South East Asia	1990	2008	121,272‡	49,344	71,928§	7.6	10.0	2.4

The Mega Rice Project forms part of the ~155,000 km^2 of peatlands that cover Borneo (Kalimantan, Sabah and Sarawak), Sumatra and peninsular Malaysia (~60% of total peatlands in southeast Asia[16]). In 1990, approximately 50% (75,800 km^2) of this land area was classed as intact PSF, with 'minor or no sign of human activity'[16]. In 2008 it was estimated that as a result of anthropogenic peatland disturbance, only 10% (15,600 km^2) of intact PSF remained, which equates to a PSF loss of 2.15% per year[16]. The TOC fluxes used in calculations are: 62.5 $g\,C\,m^{-2}\,yr^{-1}$ for intact PSF and 96.6 $g\,C\,m^{-2}\,yr^{-1}$ for disturbed PSF. *Area data taken from ref. 40. †Area of remaining intact PSF in 1990 was 48.9% of 155,020 km^2 (ref. 16) ‡Area of remaining intact PSF in 1990 was 48.9% of 248,000 km^2 (refs 1,16). §The area of disturbed peatland (excluding industrial plantations) was calculated using actual rates of PSF loss for individual regions[3,16].

through biotic decomposition in aquatic systems[23,24] and that in other, less perturbed catchments in the humid tropics, CO_2 emitted to the atmosphere from the water surface originates from the degradation of terrestrially derived organic carbon[25]. It has also been shown that old terrestrially derived organic matter is biologically processed in both rivers and estuaries[26]. Our analysis of the relative aromaticity of the DOC, (an indicator of the recalcitrance of organic carbon within the sample as derived by specific ultraviolet absorbance; see SUVA in Table 2) suggests no significant difference in the relative lability of both young and old DOC leaching different land-cover classes. We therefore expect that much of the additional, old fluvial carbon loss will be converted to CO_2 in the aquatic system, indirectly adding to greenhouse gas emissions from the disturbed sites.

Our data show that drainage of tropical peat leads to destabilization and an ongoing collapse of its carbon store, resulting in the hitherto overlooked yet quantitatively important release of carbon via fluvial organic pathways. Our findings emphasize the need to include these fluxes in models which seek to quantify the impact of disturbance on the peatland carbon balance, and in the emission factors used by the Intergovernmental Panel on Climate Change[27]. Given that the oil-palm biofuel industry contributes to regional forest destruction, our findings highlight that it is essential to incorporate fluvial organic carbon losses within guidelines for the measurement, reporting and verification of carbon emissions under the UN REDD programme. Continuing to disregard such losses may seriously undervalue the benefits to southeast Asian nations of maintaining and restoring the peatland carbon store and sink function[28].

METHODS SUMMARY

Flux estimation. Samples for DOC and POC analysis were collected weekly in channels draining intact (PSF1) and disturbed (PSF2-3) catchments. Discharge was measured weekly for 12-week periods during the dry (June–August 2008) and wet (February–May 2009) seasons, and fortnightly at other times. For weeks without discharge measurement, this was estimated from observed relationships between measured discharge and rainfall. Catchment areas were defined using elevation data from the Shuttle Radar Topography Mission (http://www2.jpl.nasa.gov/srtm/), supported by field surveys and adjusted for the presence of artificial drainage systems. Weekly flow and carbon flux estimates were summed to produce annual values, and divided by catchment area to obtain annual specific discharge ($m\,yr^{-1}$) and fluvial carbon fluxes ($g\,C\,m^{-2}\,yr^{-1}$). We compared specific discharges against rainfall measurements, satellite monitoring data, literature values and simulations using the JULES land-surface model to confirm that calculated evapotranspiration rates were within realistic ranges for each land class (see Methods).

Radiocarbon analysis. Samples for ^{14}C analysis of DOC were collected from Borneo sites during the 2008 dry season and 2011 wet season, and exploratory samples from Malaysia during the 2008 dry season. Samples were analysed by accelerator mass spectrometry at the Scottish Universities Environmental Research Centre, East Kilbride, UK. ^{14}C results were normalized to a $\delta^{13}C$ value of $-25‰$ and expressed as '%modern' and conventional radiocarbon years (relative to a 1950 baseline; see Supplementary Figure 2). Because $DO^{14}C$ levels in water samples represent the composite signal obtained by mixing organic matter from a range of ages, we modelled the age distribution of DOC by fitting a simple model of DOC production to observations, assuming an exponentially declining input of DOC with increasing peat depth (see Supplementary Information).

Full Methods and any associated references are available in the online version of the paper.

Received 24 July; accepted 21 November 2012.

1. Page, S. E., Rieley, J. O. & Banks, C. J. Global and regional importance of the tropical peatland carbon pool. *Glob. Change Biol.* **17,** 798–818 (2011).
2. Page, S. E. *et al.* The amount of carbon released from peat and forest fires in Indonesia during 1997. *Nature* **420,** 61–65 (2002).
3. Hooijer, A. *et al.* Current and future CO_2 emissions from drained peatlands in Southeast Asia. *Biogeosciences* **7,** 1505–1514 (2010).
4. Gorham, E. Northern peatlands: role in the carbon cycle and probable responses to climatic warming. *Ecol. Appl.* **1,** 182–195 (1991).
5. McDowell, W. H. & Likens, G. E. Origin, composition and flux of dissolved organic carbon in the Hubbard Brook valley. *Ecol. Monogr.* **58,** 177–195 (1988).
6. Michalzik, B., Kalbitz, K., Park, J. H., Solinger, S. & Matzner, E. Fluxes and concentrations of dissolved organic carbon and nitrogen—a synthesis for temperate forests. *Biogeochemistry* **52,** 173–205 (2001).
7. Mulholland, P. J. & Kuenzler, P. J. Organic carbon export from upland and forested wetland watersheds. *Limnol. Oceanogr.* **24,** 960–966 (1979).
8. Page, S. E, *et al.* Ecological restoration of tropical peatlands in Southeast Asia. *Ecosystems* **12,** 888–905 (2009).
9. Hoscilo, A., Page, S. E., Tansey, K. J. & Rieley, J. O. Effect of repeated fires on land-cover change on peatland in southern Central Kalimantan, Indonesia, from 1973 to 2005. *Int. J. Wildland Fire* **20,** 578–588 (2011).
10. Limpens, J. *et al.* Peatlands and the carbon cycle: from local processes to global implications—a synthesis. *Biogeosciences* **5,** 1475–1491 (2008).
11. Schiff, S. L. *et al.* Export of DOC from forested catchments on the Precambrian Shield of central Ontario: Clues from ^{13}C and ^{14}C. *Biogeochemistry* **36,** 43–65 (1997).
12. Benner, R., Benitez-Nelson, B., Kaiser, K. & Amon, R. M. W. Export of young terrigenous dissolved organic carbon from rivers to the Arctic Ocean. *Geophys. Res. Lett.* **31,** L05305 (2004).
13. Palmer, S. M. *et al.* Sources of organic and inorganic carbon in a headwater stream: evidence from carbon isotope studies. *Biogeochemistry* **52,** 321–338 (2001).
14. Evans, C. D. *et al.* Evidence against recent climate-induced destabilisation of soil carbon from ^{14}C analysis of riverine dissolved organic matter. *Geophys. Res. Lett.* **34,** L07407 (2007).
15. Raymond, P. A. *et al.* Flux and age of dissolved organic carbon exported to the Arctic Ocean: a carbon isotopic study of the five largest arctic rivers. *Glob. Biogeochem. Cycles* **21,** GB4011 (2007).
16. Miettinen, J. & Liew, S. C. Degradation and development of peatlands in peninsular Malaysia and in the islands of Sumatra and Borneo since 1990. *Land Degrad. Dev.* **21,** 285–296 (2010).
17. Page, S. E. *et al.* A record of late Pleistocene and Holocene carbon accumulation and climate change from an equatorial peat bog (Kalimantan, Indonesia): implications for past, present and future carbon dynamics. *J. Quat. Sci.* **19,** 625–635 (2004).
18. Hirano, T. *et al.* Carbon dioxide balance of a tropical peat swamp forest in Kalimantan, Indonesia. *Glob. Change Biol.* **13,** 412–425 (2007).
19. Moore, S., Gauci, V., Evans, C. D. & Page, S. E. Fluvial organic carbon losses from a Bornean blackwater river. *Biogeosciences* **8,** 901–909 (2011).
20. Montanarella, L., Jones, R. J. A. & Hiederer, R. The distribution of peatland in Europe. *Mires Peat* **1,** 1–10 (2006).
21. Billett, M. F. *et al.* Carbon balance of UK peatlands: current state of knowledge and future research challenges. *Clim. Res.* **45,** 13–29 (2010).
22. Nilsson, M. *et al.* Contemporary carbon accumulation in a boreal oligotrophic minerogenic mire - a significant sink after accounting for all C-fluxes. *Glob. Change Biol.* **14,** 2317–2332 (2008).
23. Battin, T. J. *et al.* The boundless carbon cycle. *Nature Geosci.* **2,** 598–600 (2009).
24. Cole, J. J. *et al.* Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10,** 172–185 (2007).
25. Mayorga, E. *et al.* Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. *Nature* **436,** 538–541 (2005).
26. Caraco, N., Bauer, J. E., Cole, J. J., Petsch, S. & Raymond, P. Millennial-aged organic carbon subsidies to a modern river food web. *Ecology* **91,** 2385–2393 (2010).
27. Intergovernmental Panel on Climate Change (IPCC) *Good Practice Guidance for Land Use, Land-Use Change and Forestry (LULUCF)* (eds Penman, J. *et al.*) Sections 3.2, 3.5 (IPCC National Greenhouse Gas Inventories Programme, Technical Support Unit, 2003); http://www.ipcc-nggip.iges.or.jp/public/gpglulucf/gpglulucf_files/GPG_LULUCF_FULL.pdf.
28. Murdiyarso, D., Hergoualc'h, K. & Verchot, L. V. Opportunities for reducing greenhouse gas emissions in tropical peatlands. *Proc. Natl Acad. Sci.* **107,** 19655–19660 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements S.M. was supported by a NERC PhD studentship (NE/F008813/1). Radiocarbon analyses were supported by the Natural Environment Research Council (NERC) and the Open University (CEPSAR) via the NERC Radiocarbon Facility (Environment), Allocations 1323.1008 and 15.91. A.J.W. was supported by the Joint DECC/Defra Met Office Hadley Centre Programme (GA01101). A. Hoscilo provided land-cover change estimates. We thank I. Mohammed, K. Kusin and L. Graham for field assistance. The Malaysian work was supported by the Royal Society and British Council and we thank P. Kuppan, N. Willis and F. Md. Yussof for field support.

Author Contributions V.G., S.E.P. and C.D.E. conceived and led the research conducted in Kalimantan. S.M., V.G., S.E.P. and C.D.E. designed the study and S.M. performed all the Kalimantan field data collection and analysis. C.D.E. and M.H.G. coordinated, analysed and interpreted the radiocarbon component of the work. S.M., V.G. and S.E.P. performed the scaling-up calculations. C.F. conceived and led the Malaysian study, T.G.J. performed the field data collection and analysis, A.H. provided hydrological data and interpreted land surface information to allow catchment definition. A.J.W. provided modelled estimates of evapotranspiration. S.H.L. provided expertise on the history of land-cover change and field site selection. S.M., V.G., S.E.P. and C.D.E. led the writing of the paper. All authors discussed results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.G. (v.gauci@open.ac.uk).

METHODS

Site descriptions. PSF1 is situated in the Sebangau National Park to the west of the Sebangau river and consists of a continuum of forest types from the river (riverine forest) to the centre of the peat dome (tall interior forest)[29]. The peat dome that makes up PSF1 ranges in thickness from less than a metre at the edge to 12.6 m in the centre, averaging 7.8 m over the whole dome[30]. The water table is above the surface for most of the year and reaches a maximum depth of 40 cm during the dry season in dry years[31]. PSF2 was deforested, moderately drained (3–6 m wide, 2–3 m deep channels, peat thickness <1–5 m) and had been subject to three fire events before the study and since Mega Rice Project implementation (in 1997, 2002 and 2006). PSF3 was also deforested, but subject to more intense drainage (15–25 m wide, 4–7 m deep channels, peat thickness <1–8 m) and had burned on two occasions before the study (in 1997 and 2006). The vegetation at PSF2 and PSF3 is dominated by ferns with limited woody regrowth. Stream flow in all study sites is unregulated. Water tables fluctuate according to rainfall but as a consequence of artificial drainage in PSF2 and PSF3, they remain below the peat surface for most of the year, reaching maximum depths of 140 cm in dry years[8]. The effect of such extreme drainage includes high rates of subsidence, indicative of aerobic decomposition of the peat, concentrated in the first few hundred metres from drainage channels. This is not the case in drained industrial plantations where, under usual practice, the water table is regulated at the most favourable depths for crop growth (ideally 60–80 cm for oil palm and acacia but often deeper).

Water chemistry sampling and measurement. DOC and POC samples were collected and discharge measurements taken at weekly intervals for two 12-week periods during the peak of the dry (June–August 2008) and wet (February–May 2009) seasons and fortnightly for the remaining weeks in the year, totalling 38 weeks. 2008 was unexceptional with respect to fire incidences and climate, being neither an El Niňo nor La Niňa year. For the remaining weeks, samples were collected and discharge data were inferred from rainfall data (via catchment-specific relationships between weekly rainfall data and discharge data). Five replicate flow rates and water samples were collected from each catchment outlet, representing the drainage channel cross-sectional area. Samples were collected in pre-rinsed 60 ml Nalgene bottles and water temperature, pH and electrical conductivity were recorded immediately after collection using portable pH (Hanna HI9024D) and electrical conductivity (Hanna HI8633) meters.

To derive the POC concentration, sampled river water was filtered using pre-rinsed 0.45-μm cellulose acetate membrane filters (Whatman) under partial vacuum (Mityvac, Nalgene). The residue and filter were retained and oven-dried (24 h at 40 °C) to quantify particulate matter, assumed to be equal to particulate organic matter given the dominance of peat soil in the catchment. Particulate organic matter was converted to POC assuming a 50% carbon content[32]. Filtrate was acidified to pH 2.0 with dilute sulphuric acid (20%), stored at around 2 °C and analysed upon return to the UK. DOC was determined using a Total Organic Carboniser (Shimadzu, TOC-V_{CPN}) following the non-purgeable organic carbon method. DOC/POC concentrations were then combined with discharge rates to calculate the TOC flux from each of the catchments. Ultraviolet–visible absorbance measurements were performed on a Molecular Sciences plate reader (model M2e) and a Milli-Q blank reading was taken to subtract from each sample. A quartz cell with 1.0-cm path length was used.

Discharge measurement, hydrology and flux calculation. The cross-sectional area (A, in m^2) was measured and five replicate flow rate (F, in $m\,s^{-1}$) measurements using a handheld impeller flow meter were also taken, using $F = 0.000854C + 0.05$, where C is impeller counts per minute. Hence we could calculate the discharge (Q, in $m^3\,s^{-1}$) from each channel using $Q = F \times A$. Precision for this method was better than ±5%. Weekly TOC fluxes were estimated by multiplying TOC concentration by discharge for each catchment, which was divided by the total catchment area. For each of the three study catchments, areas were estimated on the basis of the limited data available on elevation and peat depth, derived from the Shuttle Radar Topography Mission (SRTM; http://www2.jpl.nasa.gov/srtm/) 90 data and field surveys. These areas were then refined using field observations of artificial drainage systems, which dominate discharge patterns in the disturbed PSF sites. This enabled us to make estimates of discharges and carbon fluxes at each measurement time point, which we then annualized. We evaluated whether these specific discharges, in combination with rainfall rates as determined from field measurements and "Tropical Rainfall Measuring Mission" satellite monitoring[33], yielded acceptable evapotranspiration values as judged against literature values and where no data exist (that is, for deforested disturbed PSF type catchments), simulations in the JULES[34,35] land surface model.

Evapotranspiration for each land-cover class was inferred as the difference between rainfall and the sum of discharge. Inferred evapotranspiration rates are estimated as PSF1 = 1,903 mm yr^{-1}; PSF2 = 1,066 mm yr^{-1} and PSF3 = 1,086 mm yr^{-1}. For tropical lowland forest with rainfall over 2,000 mm yr^{-1}, worldwide values of between 1,200 to 1,800 mm yr^{-1} are reported[37] and for high rainfall sites such as ours (that is, >2,500–2,700 mm annual rainfall), evapotranspiration rates as high as 2,180 mm and 2,420 mm have been recorded[38,39]. It is thought that at such high rainfall rates, canopy interception and potential evaporation take on far greater importance than is currently represented in models[40], causing such models to underestimate evapotranspiration. After forest clearing, evapotranspiration has been shown generally to decrease[37], but there are no known measurements of evapotranspiration for deforested areas of the Mega Rice Project with which to validate our estimates. We therefore performed simulations of the effect of forest clearance on evapotranspiration for the Palangkaraya region using the Joint UK Land Environment Simulator JULES version 3.0 (refs 34, 35). The model was parameterized to represent PSF2 and PSF3 as combinations of grasses with C3 and C4 photosynthetic pathways and bare soil. The model was spun-up by looping over 1950–1970 until soil moisture stores stabilized and then run between 1970–2000 to derive 30-year climatology values of evapotranspiration for our two disturbed PSF sites. The model does not simulate the effects of drainage and is parameterized using ancillary information on soil properties taken from the Harmonised World Soils Database[36].

The model was forced by extracting a single half-degree gridbox of meteorological forcing data from the WATCH 20th Century forcing data set[41]. In addition, the model was forced with seasonally varying leaf area index for vegetated surface types. These were derived from climatological values of leaf area index from MODIS 1-km remote sensing data (http://modis.gsfc.nasa.gov/) classified using International Geosphere-Biosphere Programme (IGBP) land-cover classes and aggregated to the half-degree scale. Results of the simulations yielded evapotranspiration rates of 799 mm yr^{-1} for bare soil and ~1,150 mm yr^{-1} for C3/C4 grasses, which are consistent with our inferred evapotranspiration rates using the stated measurement approach, given that much of our catchments are sparsely vegetated bare soil. Finally, we checked our estimates of evapotranspiration for the PSF1 and disturbed PSF areas against measured concentrations of chloride (Cl^-, measured by ion chromatography) for samples collected across all sites. The chloride balance approach to evapotranspiration estimation assumes that all runoff Cl^- is derived from atmospheric deposition, that these inputs are consistent across the sites, and that it is unreactive during transport through the catchment (for example, ref. 42). Based on our water-balance estimates, we predicted that Cl^- concentrations in the disturbed PSF channels should be 53% of those from the PSF1 areas, owing to reduced evaporative concentration. The measured value was 43%, agreeing fairly well with this prediction, and suggesting that, if anything, evapotranspiration reduction between the PSF1 and disturbed PSF might be greater than our hydrological measurements suggest. In this case, our calculated increases in DOC flux following peatland drainage and deforestation would represent conservative estimates.

29. Boehm, H.-D. V. & Siegert, F. Ecological impact of the one million hectare rice project in central Kalimantan, Indonesia using remote sensing and GIS. *22nd Asian Conf. Remote Sensing* 1, 439–444, EME-08 (CRISP, 2001); http://www.crisp.nus.edu.sg/~acrs2001/pdf/126boehm.pdf.
30. Page, S. E., Rieley, J. O., Shotyk, Ø. W. & Weiss, D. Interdependence of peat and vegetation in a tropical peat swamp forest. *Phil. Trans. R. Soc. B* **35,** 1885–1897 (1999).
31. Takahashi, H., Usup, A., Hayasaka, H. & Limin, S. H. in *Proceedings of the International Symposium on Land Management and Biodiversity in Southeast Asia, Bali, Indonesia, 17–20 September 2002* (eds Osaki, M. *et al.*) 311–314 (Hokkaido University and Indonesian Institute of Sciences, 2003).
32. Hope, D., Billett, M. F. & Cresser, M. S. A review of the export of carbon in river water: fluxes and processes. *Environ. Pollut.* **84,** 301–324 (1994).
33. Vernimmen, R. R. E., Hooijer, A., Mamenun, E. A. & van Dijk, A. I. J. M. Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia. *Hydrol. Earth Syst. Sci.* **16,** 133–146 (2012).
34. Clark, D. B. *et al.* The Joint UK Land Environment Simulator (JULES), model description—Part 2: Carbon fluxes and vegetation dynamics. *Geosci. Model Dev.* **4,** 701–722 (2011).
35. Best, M. J. *et al.* The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.* **4,** 677–699 (2011).
36. FAO/IIASA/ISRIC/ISS-CAS/JRC *Harmonized World Soil Database (version 1.1)* (Food and Agriculture Organisation and International Iinstitute for Applied Systems Analysis, 2009); http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/.
37. Bruijnzeel, L. A. *Hydrology of Moist Tropical Forests and Effects of Conversion: a State of Knowledge Review* (UNESCO International Hydrological Programme, 1990); http://unesdoc.unesco.org/images/0009/000974/097405eo.pdf.
38. Schellekens, J., Bruijnzeel, L. A., Scatena, F. N., Bink, N. J. & Holwerda, F. Evaporation from a tropical rain forest, Luquillo Experimental Forest, eastern Puerto Rico. *Wat. Resour. Res.* **36,** 2183–2196 (2000).
39. Kume, T. *et al.* Ten-year evapotranspiration estimates in a Bornean tropical rainforest. *Agric. For. Meteorol.* **151,** 1183–1192 (2011).
40. Komatsu, H., Cho, J., Matsumoto, K. & Otsuki, K. Simple modeling of the global variation in annual forest evapotranspiration. *J. Hydrol.* **420/421,** 380–390 (2012).
41. Weedon, G. P. *et al.* Creation of the WATCH Forcing Data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *J. Hydrometeorol.* **12,** 823–848 (2011).
42. Hedin, L. O. *et al.* Patterns of nutrient loss from unpolluted, old-growth temperate forests—evaluation of biogeochemical theory. *Ecology* **76,** 493–509 (1995).

doi:10.1038/nature11832

A Y-like social chromosome causes alternative colony organization in fire ants

John Wang[1,2]*, Yannick Wurm[1,3,4]*, Mingkwan Nipitwattanaphon[1], Oksana Riba-Grognuz[1,4], Yu-Ching Huang[2], DeWayne Shoemaker[5] & Laurent Keller[1]

Intraspecific variability in social organization is common, yet the underlying causes are rarely known[1–3]. In the fire ant *Solenopsis invicta*, the existence of two divergent forms of social organization is under the control of a single Mendelian genomic element marked by two variants of an odorant-binding protein gene[4–8]. Here we characterize the genomic region responsible for this important social polymorphism, and show that it is part of a pair of heteromorphic chromosomes that have many of the key properties of sex chromosomes. The two variants, hereafter referred to as the social B and social b (SB and Sb) chromosomes, are characterized by a large region of approximately 13 megabases (55% of the chromosome) in which recombination is completely suppressed between SB and Sb. Recombination seems to occur normally between the SB chromosomes but not between Sb chromosomes because Sb/Sb individuals are non-viable. Genomic comparisons revealed limited differentiation between SB and Sb, and the vast majority of the 616 genes identified in the non-recombining region are present in the two variants. The lack of recombination over more than half of the two heteromorphic social chromosomes can be explained by at least one large inversion of around 9 megabases, and this absence of recombination has led to the accumulation of deleterious mutations, including repetitive elements in the non-recombining region of Sb compared with the homologous region of SB. Importantly, most of the genes with demonstrated expression differences between individuals of the two social forms reside in the non-recombining region. These findings highlight how genomic rearrangements can maintain divergent adaptive social phenotypes involving many genes acting together by locally limiting recombination.

The fire ant *S. invicta* is characterized by a remarkable form of social polymorphism under the control of a single Mendelian factor[4–6,8]. Remarkably, a genomic element marked by the gene *Gp-9* determines whether workers tolerate a single fertile queen (monogyne social form) or several fertile queens (polygyne social form) in their colony. Colonies containing only homozygous *Gp-9BB* workers accept only a single *Gp-9BB* queen, whereas colonies containing both *Gp-9BB* and *Gp-9Bb* workers will invariably accept several queens, but only *Gp-9Bb* queens[7,8]. The monogyne and polygyne social forms differ in numerous important aspects of their biology, including the level of aggression between colonies and how new colonies are initiated[9]. These important behavioural differences are associated with a suite of morphological and life-history differences among individuals with alternative *Gp-9* genotypes, including queen fecundity, their tendency to accumulate fat during sexual maturation, the odour of mature queens, the number of sperm produced by males, and the size of workers[4–7,10–12].

The fact that *Gp-9* codes for an odorant-binding protein (OBP), coupled with evidence that selection acted to drive the molecular divergence of *Gp-9b* alleles from the ancestral *Gp-9B* allele, has led to the suggestion that this gene may have a direct role in regulating social organization by means of chemical communication[7,13]. However, because it is unlikely that an OBP also affects traits as diverse as female size, female fecundity and male spermatogenesis, it has been suggested that *Gp-9* might be part of a supergene comprising many genes in tight linkage[10,12,14]. The existence of such clusters of loci facilitating the co-segregation of adaptive variation has been demonstrated in some classical cases of floral types (for example, ref. 15) and insect mimicry (for example, see refs 16, 17), but it remains an open question whether a polymorphic supergene is responsible for controlling the important phenotypic differences and the large suite of traits associated with variation in social organization of *S. invicta* colonies.

We used restriction-site-associated DNA (RAD) tag sequencing[18] to investigate whether *Gp-9* is located within a region with reduced recombination (supergene). In a first experiment, we obtained a total of 121 million RAD tag sequences from 95 *Gp-9B* haploid sons of a *Gp-9BB* monogyne queen (M013, Supplementary Table 1). We retained 4,983 high-coverage biallelic RAD markers from 87 informative males passing our filtering criteria to generate a genetic linkage map. The results, together with those from two other independent monogyne families, revealed 16 main linkage groups (Supplementary Figs 1–3) corresponding to the 16 chromosomes of *S. invicta*[19].

In a second experiment, we generated 92 million RAD tag sequences from 110 haploid sons of a *Gp-9Bb* polygyne queen (P034, Supplementary Table 1). Ninety-two informative males (45 *Gp-9B* and 47 *Gp-9b*) were used to construct a genetic linkage map comprising 2,796 biallelic RAD markers. Fifteen of the sixteen linkage groups in this new map were largely co-linear to those identified for the genetic maps above based on common RAD markers and their associated physical scaffolds. However, the linkage group containing *Gp-9* differed markedly between the two experiments, with 285 (10.2%) markers exhibiting no recombination with *Gp-9* in the sons of the *Gp-9Bb* queen (Supplementary Fig. 4). This non-recombining region corresponds to approximately 13.8 megabases (Mb) of the estimated 23 Mb assembled linkage group of this chromosome based on the sizes of the physical scaffolds identified by the genetic markers for this family and the three monogyne families.

Further RAD sequencing (RADseq) data from the male offspring of three *Gp-9Bb* polygyne queens ($n = 31$ (P008), 46 (P016) and 46 (P033) sons; Supplementary Figs 5–7 and Supplementary Table 1) confirmed that the non-recombining supergene marked by *Gp-9b* is a general characteristic of *S. invicta*, with a complete absence of recombination over 13.2% (478 out of 3,614), 12.1% (160 out of 1,380) and 11.0% (812 out of 7,360) of the markers assayed in the respective three families. There was very strong overlap among the non-recombining scaffolds in the *Gp-9*-linked region of family P034 and these three families (P008, 31 out of 31; P016, 19 out of 21; P033, 41 out of 41). Overall, the estimated size of the non-recombining region is 12.7 Mb based on the combined data of these four polygyne families (Fig. 1a),

[1]Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland. [2]Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan. [3]School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK. [4]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland. [5]USDA-ARS Center for Medical, Agricultural, and Veterinary Entomology, 1600/1700 Southwest 23rd Drive, Gainesville, Florida 32608, USA.
*These authors contributed equally to this work.

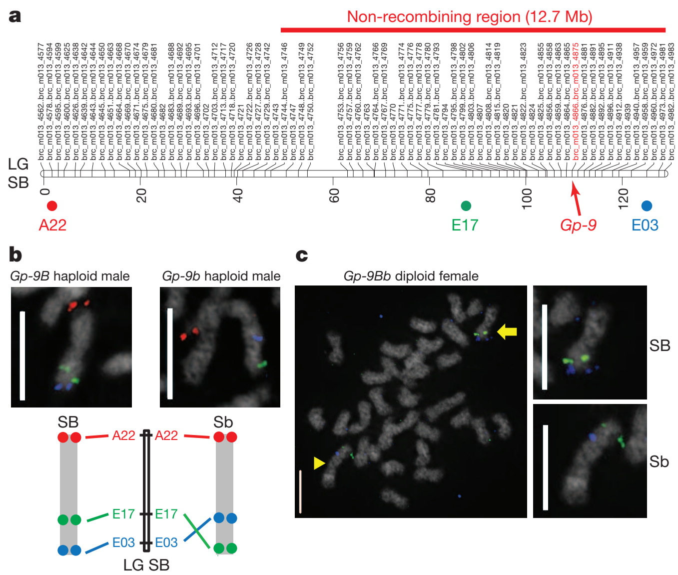


Figure 1 | Fine-scale mapping and BAC-FISH analysis of the social chromosome. **a**, Fine-scale linkage map of the SB social chromosome derived from RADseq analysis of male offspring from a monogyne *Gp-9BB* queen (M013). Genetic positions in centimorgans (below) and RAD marker names (above) are indicated. Each genetic marker has a prefix (brc_m013_) and a number based on its serial position on the map. Multiple markers that have the same map position are indicated using 'first…last' marker notation. The positions of *Gp-9*, BAC probes and the extent of the non-recombining region in *Gp-9Bb* queens are indicated. Genetic maps for all seven families are in Supplementary Figs 1–7 and precise genetic and physical positions of genetic markers are in Supplementary Tables 8–14. LG, linkage group. **b**, BAC-FISH identifies an inversion between the SB and Sb social chromosomes. Social chromosomes from a *Gp-9B* haploid male (SB chromosome, left) and *Gp-9b* haploid male (Sb chromosome, right). The bottom panel shows a schematic interpretation of hybridization patterns as well as BAC positions on the SB genetic map. Images with all chromosomes of the respective cells are in Supplementary Fig. 8. **c**, BAC-FISH on full chromosome complement from one cell of a diploid *Gp-9Bb* female (SB/Sb chromosomes). Both orientations corresponding to SB (arrow) and Sb (arrowhead) can be observed. Magnified views of respective social chromosomes are on right. Chromosomes are counterstained with 4′,6-diamidino-2-phenylindole (DAPI; white) and hybridized with fluorescently labelled BAC probes: A22 (red, **b**), E17 (green, **b**, **c**) and E03 (blue, **b**, **c**). Scale bars, 5 µm.

and includes at least 616 (3.7%) of the 16,522 known *S. invicta* genes[20]. This number is probably an underestimate because only 82% (13,557 out of 16,522) of the genes have been assigned a genetic map position. Importantly, there was no evidence of reduced recombination in this 12.7-Mb region in the three monogyne families headed by *Gp-9BB* queens (two-factor mixed-model analysis of variance (ANOVA) on log recombination rate, $F_{1,44} = 0.17$, $P = 0.68$).

The suppressed recombination near *Gp-9* was further confirmed by examining insertion-deletion (indel) alleles in four *Gp-9B* and four *Gp-9b* males from each of five further independent polygyne families. We sequenced and performed a *de novo* assembly of the genome of a *Gp-9b* male as previously done for the *Gp-9B* genome to identify indel polymorphisms in the *Gp-9*-linked region and also to permit detailed sequence comparisons (below) between the *Gp-9B* and *Gp-9b* alleles in the non-recombining supergene[20]. Indel-specific PCR assays for 16 randomly selected *Gp-9*-linked loci confirmed a complete lack of recombination between the two *Gp-9*-linked regions (Supplementary Table 2). Having established that the *B* and *b* variants of this linkage group are a general characteristic of *S. invicta*, we hereafter refer to them as social B and social b (SB and Sb) chromosomes.

Chromosomal rearrangements are a key mechanism preventing recombination between homologous chromosomes[17,21]. We performed bacterial artificial chromosome fluorescent *in situ* hybridization (BAC-FISH) to test whether an inversion between SB and Sb could contribute to reduced recombination. Multicolour BAC-FISH on *Gp-9B* haploid males revealed a BAC hybridization pattern concordant with the SB genetic map in the five individuals analysed (Fig. 1b and Supplementary Figs 8 and 9a–d), whereas all six *Gp-9b* haploid males analysed exhibited an inverted BAC hybridization pattern for the probes in the non-recombining region (Fig. 1b and Supplementary Figs 8 and 9a–d). Both orientations were observed in the three *Gp-9Bb* diploid females analysed (Fig. 1c and Supplementary Fig. 9e). These data demonstrate the presence of a large inversion, which, on the basis of the map positions of the BACs tested, is at least 9.3 Mb.

In addition to this large chromosomal inversion, sequence comparisons between SB and Sb revealed a smaller 48 kilobase (kb) inversion in the non-recombining region (Supplementary Fig. 10). Intriguingly, this inversion is adjacent to a 3-kb transposed sequence that interrupts SI2.2.0_02248, a putative acyl-CoA desaturase. Genes in this family have a role in pheromone and cuticular hydrocarbon synthesis[22], raising the possibility that this rearrangement could be involved in odour differences existing between *Gp-9BB* and *Gp-9Bb* queens, which are the cues known to be used by workers to discriminate between queens of alternative genotypes[7]. High-throughput RNA sequencing (RNA-seq) comparisons did show that this gene was expressed at a significantly lower level in *Gp-9b* males (2.9 reads per kilobase of exon model per million mapped reads (RPKM)) than in *Gp-9B* males (10.7 RPKM; likelihood ratio test on negative binomial generalized linear model $P = 0.001$). Moreover, the analysis of males from the five additional families above confirmed that this rearrangement is a general characteristic of the SB and Sb chromosomes (all 20 *Gp-9B* males had PCR product sizes specific to SB, whereas all 20 *Gp-9b* males had PCR product sizes specific to Sb; Supplementary Table 2). Thus, this rearrangement constitutes a general difference between the SB and Sb chromosomes and may be directly involved in odour differences between *Gp-9BB* and *Gp-9Bb* queens.

The selective pressures acting on the two heteromorphic social chromosomes identified in *S. invicta* should be similar to those acting on sex chromosomes and other genomic regions experiencing epistatic selection for reduced recombination[23]. Y chromosomes are generally thought to have evolved from an autosome after suppression of recombination between a 'sex locus', which makes bearers more likely to develop into a male than a female, and a nearby gene with sexually antagonistic alleles. Over evolutionary time, cessation of recombination occurs between the X and Y chromosomes, allowing additional genes beneficial for males but harmful to females to accumulate on the Y chromosome[21,24]. Analogous to the Y chromosome, which is only found in males (or the W chromosome in females), the Sb chromosome only occurs in one of the two alternative social forms (the polygyne form) of *S. invicta*. Colony queen number was likely to be a plastic and non-genetic ancestral trait in fire ants, as is probably the case in many other ant species[1,9]. Thus, in a similar manner to Y chromosome evolution, genes conferring both a higher fitness in polygyne colonies and a higher probability of queens of joining such colonies should be selected to become linked in a non-recombining region.

Importantly, the available data suggest that most phenotypic differences between individuals of the two social forms are directly associated with differences between the non-recombining regions of the SB and Sb chromosomes. Nineteen of the 27 (70%) genes previously shown to be differentially expressed between *Gp-9BB* and *Gp-9Bb* workers[25] that could be mapped to linkage groups were in the non-recombining region (Fig. 2), although this region only includes 4.9% of the 7,282 genes on the microarray with a genetic map position (hypergeometric test, $P < 10^{-22}$). Microarray experiments revealed similar results for genes differentially expressed between queens and males of alternative *Gp-9* genotype. Of the 38 genes differentially expressed between 1-day-old virgin queens with genotypes *Gp-9BB* and *Gp-9Bb*, 15 could be mapped to known linkage groups, four (27%) of which were located in the non-recombining region (Fig. 2 and Supplementary

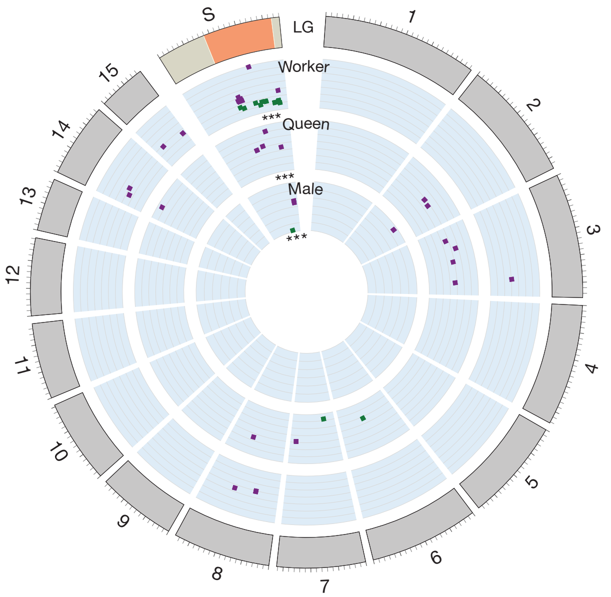


Figure 2 | Expression of genes associated with *Gp-9* genotype are overrepresented on the social chromosome. The outer circle depicts the *S. invicta* chromosome ideograms. The social chromosome (S) is subdivided into the non-recombining (orange) and recombining (tan) regions. Genes differentially expressed between individuals of alternative *Gp-9* genotypes in mature adult workers[25], young adult queens and male pupae are plotted as squares according to their genomic location. The relative expression level ($\log_2$-transformed) of *Gp-9BB* to *Gp-9Bb* (worker and queens) or *Gp-9B* to *Gp-9b* (male) is also indicated, with squares closer to the circle centre having greater expression in *Gp-9BB* or *Gp-9B* individuals. Colours highlight direction of gene expression: green, expression in *Gp-9BB* (or *Gp-9B*) > *Gp-9Bb* (or *Gp-9b*); and purple, reversed. ***$P < 0.001$, hypergeometric test.

Table 3; hypergeometric test, $P < 0.0008$). Similarly, of the five genes that were differentially expressed between *Gp-9B* and *Gp-9b* male pupae, four could be mapped to known linkage groups, three (75%) of which were located in the non-recombining region (Fig. 2 and Supplementary Table 3; hypergeometric test, $P < 0.00012$). Finally, 15 of the 616 genes in the non-recombining regions had d_N/d_S values (that is, the ratio of the number of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site) significantly greater than one (Supplementary Table 4). High d_N/d_S values may be expected under adaptive or relaxed purifying selection. Six of the eight genes for which expression data are available exhibited a male-biased expression pattern (Supplementary Table 4), suggesting that they are functional and exposed to selection in males. Thus, the high d_N/d_S values of some genes may reflect adaptive evolution[26] driving phenotypic differences between individuals of the two social forms.

Another similarity between the Sb chromosome and the Y (or W) sex chromosome is that Sb contains one or more (recessive-lethal) deleterious mutations. Similar to Y chromosomes and supergenes in which there has been selection for suppressed recombination, the non-recombining region of Sb is expected to experience less efficient purifying selection, and thus increased accumulation of mildly deleterious mutations. The non-viability of Sb/Sb individuals is one clear example and effectively prevents recombination in the 12.7-Mb region. Such non-recombining regions also feature higher frequencies of repetitive elements, longer introns and increased fixation of non-synonymous substitutions[21,23,24,27]. Consistent with these predictions, whole-genome sequence analysis of one *Gp-9B* son and one *Gp-9b* son from each of seven unrelated *Gp-9Bb* queens revealed that 17% (237 out of 1,461) of the known fire ant repetitive elements were significantly more frequent in *Gp-9b* than *Gp-9B* males, whereas only 0.7% (9 out of 1,461) were more frequent in *Gp-9B* males (false discovery rate (FDR)-corrected two-sided paired *t*-tests on numbers of reads matching each repetitive element; Supplementary Table 5). Also, scaffolds in the non-recombining region of Sb were 2.7 times shorter than those of SB, and 3.3 times shorter than scaffolds from the rest of the *Gp-9b* assembly (Fig. 3), consistent with the prediction that higher frequency of repetitive elements in the non-recombining Sb region results in increased difficulties in assembly of this region. We also found that 49 out of 472 (10.4%) of the intron-containing protein-coding genes present in both genome assemblies had an intron at least 500 base pairs (bp) larger in the non-recombining region of Sb than SB (Fisher's exact test, $P < 10^{-10}$). Finally, the d_N/d_S ratios for genes in the non-recombining region of the social chromosome were significantly higher (0.21 ± 0.31, median $\pm$ standard error) than those for the remainder of the genome (0.12 ± 0.01; two-sided Wilcoxon rank-sum test, $P < 10^{-4}$).

Although the above analyses reveal notable similarities among the Sb, Y and W animal chromosomes, we predict that the rate of gene degeneration in the non-recombining Sb region compared with Y and W animal chromosomes should be relatively slow because of purifying selection acting on genes expressed in ant haploid males. A similar argument has been made for algae and bryophytes, in which sex is determined during the haploid phase of the life cycle. Generally, individuals with the U chromosome develop as females and those with the V chromosome as males. The U and V chromosomes are predicted to be partly sheltered from genetic deterioration because they are subject to purifying selection as haploids[21]. Similarly, many genes in plants are also expressed in the haploid male gametophyte and are thus under strong selection owing to competition between the X- and Y-bearing pollen during pollination[28]. By contrast, haploid Y-bearing spermatozoa in animals show very limited gene expression, which probably accounts for the higher rate of degeneration of Y-linked genes in animals than in plants[24].

Several lines of evidence suggest that purifying selection acting in haploid males may slow down degeneration of the non-recombining region of Sb. First, RNA-seq data revealed that the vast majority of

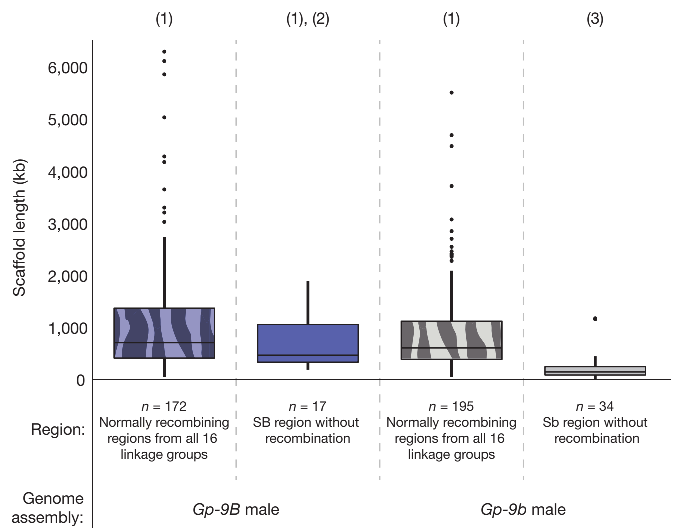


Figure 3 | Scaffolds lengths of the non-recombining region of the social chromosome (solid) and the rest of the genome (patterned) based on the genome assemblies of a *Gp-9B* (blue) and a *Gp-9b* (grey) male. Box-and-whisker plots of scaffold lengths: the top and bottom of the box are the first and third quartiles, respectively; the horizontal bar within the box is median; whiskers extend from the box to the most extreme scaffold length within 1.5× of the interquartile range of the box; data beyond the whiskers are outliers and plotted as points. Tukey's honestly significant difference (HSD) comparisons among groups marked (1) are non-significant ($P > 0.05$), the comparison between (1) and (3) is significant ($P < 10^{-7}$), and the comparison between (2) and (3) is also significant ($P < 10^{-4}$).

S. invicta genes expressed in females are also expressed in haploid males (data not shown). Second, microarray experiments between *Gp-9B* and *Gp-9b* males revealed no differentially expressed genes in adults and only five differentially expressed genes in pupae (Supplementary Table 3). Third, 99.8% of the non-gap sequences alignable between the non-recombining regions of SB and Sb were identical and largely contained the same genes. Indeed, we identified missing exons in Sb compared with SB for only five out of 616 genes (Supplementary Table 6). Finally, the sequencing of six RNA pools, each of which comprised four *Gp-9Bb* queens contributing equal quantities of RNA, showed significant allele-specific expression differences for only 10.8% (31 out of 288, Supplementary Table 7) of the genes possessing single nucleotide polymorphisms (SNPs) in the non-recombining region and no evidence of systematically higher expression of alleles on SB compared with those on Sb (binomial test, $P = 0.11$, Supplementary Fig. 11). This result contrasts with *Drosophila miranda* neo-sex chromosomes in which 80% of the genes on the neo-X chromosome have higher expression than those on the neo-Y chromosome[29]. Overall, these combined results suggest that the non-recombining region of Sb has retained the vast majority of the genes and that purifying selection has acted as a potent force to slow down their rate of degeneration.

The timing of the origin of the heteromorphic social chromosomes in fire ants is difficult to determine at present. A previous phylogenetic analysis of *Gp-9* sequences from 21 fire ant (*Solenopsis*) species, of which four (including *S. invicta*) are socially polymorphic, revealed a perfect association between queen *Gp-9* genotype and colony social form[13]. Specifically, in all the three other socially polymorphic species, all monogyne queens were homozygous for a *Gp-9B*-like allele similar to the *Gp-9B* allele found in *S. invicta*, whereas all polygyne queens had one copy of a *Gp-9b*-like allele, which again was very similar to the *Gp-9b* allele identified in *S. invicta*. Importantly, these analyses showed that all *Gp-9b*-like alleles form a monophyletic group within the *B*-like alleles, consistent with a single origin of polygyny in this group. If *Gp-9* in these three closely related species is also part of a supergene as we suspect, then recombination suppression between SB and Sb may be ancestral to speciation. Comparisons of the frequency of synonymous substitutions between SB and Sb alleles and the frequency of synonymous substitutions between genes of two leafcutter ants that diverged ~10 million years ago[30] suggest that recombination suppression in the social chromosome occurred ~390,000 years ago (Supplementary Fig. 12 and Supplementary Information), which could be less than the divergence time (K. G. Ross, unpublished data) between these four socially polymorphic *Solenopsis* species. A more precise estimate of the divergence time of SB and Sb relative to the four species is required to test whether the divergence of SB and Sb predates speciation or whether introgression of Sb across species boundaries has occurred.

In conclusion, our study shows that colony social organization in the fire ant *S. invicta* is under the control of a pair of heteromorphic social chromosomes having many of the properties of sex chromosomes. The Sb chromosome only occurs in one type of social organization, which sets the stage for a specific selection regime very similar to that acting on sex-specific Y and W chromosomes. Indeed, the Sb chromosome has many similar properties to Y and W chromosomes, including a large non-recombining region, inversions, an increased amount of repetitive elements and deleterious mutations resulting in Sb/Sb individuals being non-viable. This is the first description of a social chromosome, yet it is likely that such supergenes affecting social organization also exist in other social insects. Polymorphism in social organization has evolved independently numerous times in ants where many species have both monogyne and polygyne colonies. The occurrence of the polygyne social form is associated almost invariably with a 'polygyny syndrome' in which, as in *S. invicta*, polygyne queens are smaller, accumulate less fat during sexual maturation, have lower fecundity and initiate new colonies with the help of workers rather than independently[1,2]. It will be interesting to investigate whether these differences, as in *S. invicta*, also result from genomic rearrangements creating a supergene containing many genes, which jointly provide integrated control to maintain divergent and adaptive social phenotypes.

METHODS SUMMARY

RADseq was performed on bar-coded male offspring from monogyne and polygyne families. Individual-specific data subsets were generated from the raw sequences and then deposited into family-specific MySQL databases using the FASTX-toolkit and Perl scripts. Loci that were biallelic in a family and monoallelic (that is, never heterozygous because ant males are haploid) in individual males were used to create family-specific genetic maps. Multicolour BAC-FISH was performed on chromosome preparations from imaginal discs using social chromosome-specific clones. The newly assembled *Gp-9b* genome was compared to the *Gp-9B* reference assembly using Ruby scripts and bioinformatics tools. The d_N/d_S ratios were calculated using codeml in PAML. Microarray-based gene expression differences were assayed using custom complementary DNA microarrays and analysed using limma (Bioconductor, R). The genetic positions of differentially expressed genes were determined by blastn comparisons of the microarray cDNAs to the genetically mapped physical scaffolds. SNPs were identified by sequencing several *Gp-9B* and *Gp-9b* individuals and by RNA-seq. Allele-specific expression was assessed by RNA-seq. Complete methods and further analyses are provided in the Supplementary Information.

Received 14 June; accepted 10 December 2012.
Published online 16 January 2013.

1. Bourke, A. & Franks, N. *Social Evolution in Ants* (Princeton University Press, 1995).
2. Keller, L. Social life – the paradox of multiple-queen colonies. *Trends Ecol. Evol.* **10,** 355–360 (1995).
3. Robinson, G. E., Fernald, R. D. & Clayton, D. F. Genes and social behavior. *Science* **322,** 896–900 (2008).
4. Krieger, M. J. B. & Ross, K. G. Identification of a major gene regulating complex social behavior. *Science* **295,** 328–332 (2002).
5. Keller, L. & Ross, K. G. Phenotypic basis of reproductive success in a social insect: genetic and social determinants. *Science* **260,** 1107–1110 (1993).
6. DeHeer, C. J., Goodisman, M. A. D. & Ross, K. G. Queen dispersal strategies in the multiple-queen form of the fire ant *Solenopsis invicta*. *Am. Nat.* **153,** 660–675 (1999).
7. Keller, L. & Ross, K. G. Selfish genes: a green beard in the red fire ant. *Nature* **394,** 573–575 (1998).
8. Ross, K. G. & Keller, L. Genetic control of social organization in an ant. *Proc. Natl Acad. Sci. USA* **95,** 14232–14237 (1998).
9. Ross, K. G. & Keller, L. Ecology and evolution of social-organization: insights from fire ants and other highly eusocial insects. *Annu. Rev. Ecol. Syst.* **26,** 631–656 (1995).
10. Keller, L. & Ross, K. G. Major gene effects on phenotype and fitness: the relative roles of *Pgm-3* and *Gp-9* in introduced populations of the fire ant *Solenopsis invicta*. *J. Evol. Biol.* **12,** 672–680 (1999).
11. Keller, L. & Ross, K. G. Gene by environment interaction: effects of a single-gene and social-environment on reproductive phenotypes of Fire Ant queens. *Funct. Ecol.* **9,** 667–676 (1995).
12. Lawson, L. P., Vander Meer, R. K. & Shoemaker, D. Male reproductive fitness and queen polyandry are linked to variation in the supergene *Gp-9* in the fire ant *Solenopsis invicta*. *Proc. R. Soc. Lond. B* **279,** 3217–3222 (2012).
13. Krieger, M. J. B. & Ross, K. G. Molecular evolutionary analyses of the odorant-binding protein gene *Gp-9* in fire ants and other *Solenopsis* species. *Mol. Biol. Evol.* **22,** 2090–2103 (2005).
14. Gotzek, D. & Ross, K. G. Genetic regulation of colony social organization in fire ants: an integrative overview. *Q. Rev. Biol.* **82,** 201–226 (2007).
15. Mather, K. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution* **4,** 340–352 (1950).
16. Clarke, C. A., Sheppard, P. M. & Thornton, I. W. The genetics of the mimetic butterfly *Papilio memnon* L. *Philos. Trans. R. Soc. Lond. B* **254,** 37–89 (1968).
17. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–206 (2011).
18. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3,** e3376 (2008).
19. Glancey, B. M., Romain, M. K. S. & Crozier, R. H. Chromosome numbers of red and black imported fire ants, *Solenopsis invicta* and *Solenopsis richteri*. *Ann. Entomol. Soc. Am.* **69,** 469–470 (1976).
20. Wurm, Y. *et al.* The genome of the fire ant *Solenopsis invicta*. *Proc. Natl Acad. Sci. USA* **108,** 5679–5684 (2011).
21. Bachtrog, D. *et al.* Are all sex chromosomes created equal? *Trends Genet.* **27,** 350–357 (2011).
22. Blomquist, G. J. & Vogt, R. G. *Insect Pheromone Biochemistry and Molecular Biology: the Biosynthesis and Detection of Pheromones and Plant Volatiles* (Academic, 2003).
23. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics* (Roberts & Company, 2010).
24. Bergero, R. & Charlesworth, D. Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr. Biol.* **21,** 1470–1474 (2011).

25. Wang, J., Ross, K. G. & Keller, L. Genome-wide expression patterns and the genetic architecture of a fundamental social trait. *PLoS Genet.* **4,** e1000127 (2008).
26. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).
27. Feldman, M. W. & Liberman, U. An evolutionary reduction principle for genetic modifiers. *Proc. Natl Acad. Sci. USA* **83,** 4824–4827 (1986).
28. Correns, C. Die Rolle der männlichen Keimzellen bei der Geschlechtsbestimmung der gynodiöecischen Pflanzen. *Ber. Deut. Bot. Ges.* **26A,** 686–701 (1908).
29. Bachtrog, D. Expression profile of a degenerating neo-Y chromosome in *Drosophila. Curr. Biol.* **16,** 1694–1699 (2006).
30. Nygaard, S. *et al.* The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res.* **21,** 1339–1348 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Stoffel, C. La Mendola, N.-C. Chang, C.-Y. Kao and C.-C. Lee for helping with genotyping and molecular biology; the DEE-UNIL animal caretakers for ant husbandry; E. Johnson and P. Etter for RADseq advice; R. Nichols, J. Meunier and R. Verity for statistical advice; K. Harshman and M.-Y. Lu for Illumina sequencing support; R. Wang for FISH support; and B. Charlesworth, D. Charlesworth, H. Kaessmann, L. Ometto, J. Pannel, N. Perrin, M. Reuter, P. Reymond and K. Ross for comments. Some computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing (HPC) of the SIB Swiss Institute of Bioinformatics and the EPSRC-funded MidPlus HPC centre. This work was supported by the Biodiversity Research Center (Academia Sinica, Taiwan), Taiwan NSC grant 100-2311-B-001-015-MY3, grants from NERC and the BBSRC (BB/K004204/1), a USDA grant, several grants from the Swiss NSF and an ERC Advanced Grant.

Author Contributions J.W., Y.W. and L.K. designed the study and contributed to all stages of the project. M.N., D.D.S. and J.W. prepared samples. J.W. performed RAD sequencing, and J.W. and Y.W. performed genetic analyses. M.N. performed microarray experiments and analysed the data. O.R.-G. and Y.W. analysed RNA-seq and SNP data. J.W. and Y.W. analysed chromosomal locations of differentially expressed genes. Y.W. performed sequence assembly, genome comparisons, and molecular evolution analyses. Y.-C.H. performed FISH experiments. L.K., Y.W. and J.W. wrote the paper with input from other authors.

Author Information The microarray expression data are available at the NCBI Gene Expression Omnibus (accessions GSM1031731–GSM1031746, GSM1031779–GSM1031794, GSM1040938–GSM1040947, GSM1049807–GSM1049816 and GSM1049903–GSM1049912); sequence data are available at the NCBI Sequence Read Archive (accessions SRA061944, SRP017299, SRP017317 and SRP017322). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W. (johnwang@gate.sinica.edu.tw), Y.W. (y.wurm@qmul.ac.uk), or L.K. (laurent.keller@unil.ch).

doi:10.1038/nature11810

Genetic identification of C fibres that detect massage-like stroking of hairy skin *in vivo*

Sophia Vrontou[1], Allan M. Wong[1,2], Kristofer K. Rau[3], H. Richard Koerber[4] & David J. Anderson[1,2]

Stroking of the skin produces pleasant sensations that can occur during social interactions with conspecifics, such as grooming[1]. Despite numerous physiological studies (reviewed in ref. 2), molecularly defined sensory neurons that detect pleasant stroking of hairy skin[3,4] *in vivo* have not been reported. Previously, we identified a rare population of unmyelinated sensory neurons in mice that express the G-protein-coupled receptor MRGPRB4 (refs 5, 6). These neurons exclusively innervate hairy skin with large terminal arborizations[7] that resemble the receptive fields of C-tactile (CT) afferents in humans[8]. Unlike other molecularly defined mechanosensory C-fibre subtypes[9,10], MRGPRB4$^+$ neurons could not be detectably activated by sensory stimulation of the skin *ex vivo.* Therefore, we developed a preparation for calcium imaging in the spinal projections of these neurons during stimulation of the periphery in intact mice. Here we show that MRGPRB4$^+$ neurons are activated by massage-like stroking of hairy skin, but not by noxious punctate mechanical stimulation. By contrast, a different population of C fibres expressing MRGPRD[11] was activated by pinching but not by stroking, consistent with previous physiological and behavioural data[10,12]. Pharmacogenetic activation of *Mrgprb4*-expressing neurons in freely behaving mice promoted conditioned place preference[13], indicating that such activation is positively reinforcing and/or anxiolytic. These data open the way to understanding the function of MRGPRB4 neurons during natural behaviours, and provide a general approach to the functional characterization of genetically identified subsets of somatosensory neurons *in vivo.*

In isolated skin-nerve preparations, MRGPRB4$^+$ neurons were not electrophysiologically activated by mechanical, thermal or chemical stimuli (see Supplementary Note 1). Therefore, we sought to perform calcium imaging specifically in these neurons while stimulating the periphery of intact mice. To target genetically encoded calcium sensors to MRGPRB4$^+$ or MRGPRD$^+$ neurons, we injected neonatal *Mrgprb4-tdTomato-2A-cre* mice (Supplementary Fig. 1) or *MrgprD-EGFP-cre* mice[10] intraperitoneally (i.p.) with a Cre-dependent adeno-associated virus (AAV) expressing GCaMP3.0 (ref. 14) (Supplementary Table 1, Methods and Supplementary Note 2). A similar efficiency of viral expression ($62 \pm 3.6\%$) was observed in *MrgprD-EGFP-cre* mice (Fig. 1b, Supplementary Fig. 2a–c, g and Supplementary Note 2). This approach yielded relatively efficient expression of the genetically encoded calcium sensor in MRGPRB4::tdTomato$^+$ dorsal root ganglia (DRGs) neurons ($62 \pm 6\%$) along the rostro-caudal axis in adult mice (Fig. 1a, c, Supplementary Fig. 2d–f, h and Supplementary Note 2). Expression of GCaMP3.0 or mGCaMP3.0 was especially robust in the central spinal projections of these neurons (Fig. 1d, e). No expression of the reporter was observed in virally injected wild-type mice.

To record calcium transients in the central projections of MRGPRD$^+$ or MRGPRB4$^+$ neurons, we performed two-photon imaging through a spinal cord laminectomy while stimulating the intact animal (Fig. 1f and Supplementary Note 3). We first tested responses to centrally or peripherally applied chemical stimuli. Direct application to the spinal cord of depolarizing concentrations of KCl elicited robust fluorescence increases over baseline, $\Delta F/F$, in both MRGPRD$^+$ fibres (Fig. 1g, i, m; mean per cent increase in peak $\Delta F/F$ (MPI $\Delta F/F_{peak}$) $= 222 \pm 19\%$ ($\pm$s.e.m.); mean latency to peak (MLP) $= 8.6 \pm 3.6$ s, $n = 3$) and MRGPRB4$^+$ fibres (Fig. 1h, j, n; MPI $\Delta F/F_{peak} = 201.6 \pm 33.2\%$, MLP $= 9.3 \pm 4.15$ s, $n = 3$). We also observed responses to α,β-methylene (Me) ATP, a ligand known to activate both MRGPRB4$^+$ and MRGPRD$^+$ neurons *in vitro*[7,15], via both direct spinal application and/or peripheral injection into hairy or glabrous skin of the hindpaw, respectively (Fig. 1k–n and Supplementary Fig. 3; see also Supplementary Note 4). Finally, MRGPRB4$^+$ central fibres were activated by peripheral injection of capsaicin in mice genetically engineered to express TRPV1 in MRGPRB4$^+$ neurons, which normally do not express this channel[7] (Supplementary Fig. 4 and Supplementary Note 4). Thus, our preparation was able to detect calcium transients in both MRGPRD$^+$ and MRGPRB4$^+$ fibres by peripheral injection of specific chemical stimuli that activate these neurons.

We next tested whether this preparation could be used to image activity evoked by mechanical stimulation of the periphery. We first measured activity in MRGPRD$^+$ fibres after mechanical stimulation of the hindpaw using a custom pinching device (see Methods) (Fig. 2a). In agreement with their established role in sensing noxious punctate mechanical stimuli[10,12], MRGPRD$^+$ fibres were strongly activated by trains of pinching stimuli in the ipsilateral hindpaw (and more specifically in the particular experiment by pinching of the most distal ipsilateral digit) (Fig. 2d, e; MPI $\Delta F/F_{peak} = 77.8 \pm 8.9\%$, $n = 7$ trials per mouse). Responses were restricted to a subset of GCaMP3.0-expressing fibres within a given imaging field, whereas other fibres were unresponsive (Fig. 2c–e, h and Supplementary Fig. 5a–e). This heterogeneity probably reflects the different receptive fields of these fibres relative to the site of stimulation. Responses of MRGPRD$^+$ fibres to pinching were reproducible across trials and mice (Fig. 2d, e, h, j and Supplementary Table 2), and also specific to the ipsilateral hindpaw and to particular digits (Supplementary Fig. 6 and Supplementary Note 5).

Importantly, MRGPRD$^+$ fibres in a given region of interest (ROI) that were activated by pinching were not activated when the last digit of the ipsilateral hindpaw was stroked lightly using a brush (Fig. 2f, g, i). The same fibres could, however, be reactivated by a subsequent pinching stimulus (Supplementary Fig. 9i–l), indicating that the lack of response to brushing was not due to adaptation or desensitization produced by the pinch stimulus. These data therefore suggest a specificity of MRGPRD$^+$ fibres for punctate or focal noxious mechanical stimulation of the skin, consistent with previous physiological and behavioural studies of this subpopulation[10,12].

To functionally characterize MRGPRB4$^+$ fibres, we tested a variety of innocuous mechanical stimuli designed to simulate natural stroking or grooming, using a custom-designed brush (Fig. 3a, Supplementary Fig. 14 and Methods). Calcium transients in MRGPRB4-tdTomato$^+$ fibres (Supplementary Fig. 7) were elicited by repeated stroking (0.2–0.5 Hz)

[1]Division of Biology 156-29, California Institute of Technology, Pasadena, California 91125, USA. [2]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125, USA. [3]Department of Anatomical Sciences and Neurobiology, University of Louisville, Louisville, Kentucky 40202-1702, USA. [4]Department of Neurobiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA.

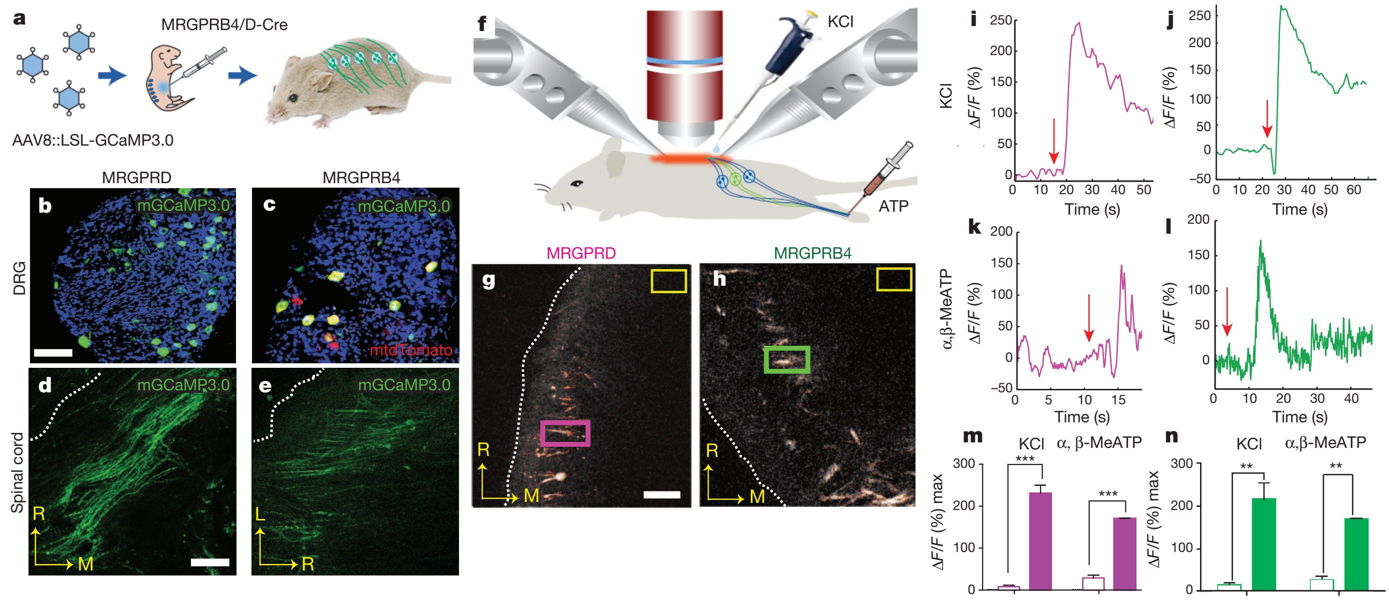


Figure 1 | *In vivo* calcium imaging in genetically defined subsets of primary sensory neurons. **a**, Schematic illustrating AAV infection. LSL, *loxP-STOP-loxP* cassette. **b**–**e**, mGCaMP3.0 expression in somata (**b**, **c**) and central afferent fibres (**d**, **e**) of MRGPRD$^+$ (**b**, **d**) or MRGPRB4$^+$ (**c**, **e**) neurons in adult mice. Dashed lines indicate lateral margin of spinal cord. Scale bars: 50 μm (**b**); 45 μm (**d**). **f**, Schematic illustrating imaging preparation; components not to scale. **g**–**n**, Calcium transients in the central projections of MRGPRD$^+$ (**g**, **i**, **k**, **m**) or MRGPRB4$^+$ (**h**, **j**, **l**, **n**) neurons, evoked by direct application of KCl to the spinal cord (**i**, **j**) or (in a different animal) peripheral injection of α,β-methylene ATP (**k**, **l**). Coloured rectangles in **g** and **h** indicate ROIs used in **i** and **j**, respectively; yellow boxes are regions for background subtraction. Scale bars: 40 μm (**g**); 20 μm (**h**). Red arrows (**i**–**l**) indicate time of stimulus delivery. **m**, **n**, Quantification of peak $\Delta F/F$ values before (open bars) versus after (filled bars) stimulation. $^{**}P < 0.01$; $^{***}P < 0.001$. All data are mean ± s.e.m. L, lateral; M, medial; R, rostral.

of relatively large areas (2–3 mm × 20–30 mm) of posterior dorsal thoracic and proximal hindlimb hairy skin (Fig. 3c–e, h; green traces), consistent with the distribution of MRGPRB4$^+$ fibres in the periphery[7]. The average forces and velocities delivered from these manual stimuli, which included a mild pressure component, were relatively dynamic but fell within the range of 20–90 mN and a speed of 0.5–2 cm s^{-1} (see Methods). As in the case of MRGPRD$^+$ neuron responses to pinching, responses in MRGPRB4$^+$ neurons to stroking were observed in particular fibres in a given field of view, were specific to the ipsilateral side of the animal and particular areas of the skin, and were reproducible across trials and animals (Fig. 3, Supplementary Figs 5, 7 and 8, Supplementary Table 3 and Supplementary Note 6). In contrast to MRGPRD$^+$ fibres, MRGPRB4$^+$ fibres were not activated by localized pinching of hairy skin in regions activated by stroking (Fig. 3f, g, i), and this selectivity was not due to desensitization (Supplementary Fig. 9a–f). These data indicate that MRGPRB4$^+$ fibres are activated by massage-like stroking of hairy skin. Thus, the two classes of cutaneous C fibres marked by expression of MRGPRB4 and MRGPRD, respectively, respond to distinct types of mechanical stimulation *in vivo*. The reason why MRGPRB4$^+$ fibres are not also activated by pinching is not clear, but could reflect their specific tuning to moving stimuli[16].

The stimuli used to activate MRGPRB4$^+$ fibres were designed to mimic stroking and allogrooming stimuli. The social interactions associated with such stimuli have been shown to be positively reinforcing in juvenile mice[17], using a conditioned place preference (CPP) assay[13,17], suggesting that these stimuli may have a positive affective valence[18]. We therefore asked whether direct activation of MRGPRB4$^+$ neurons could similarly promote a preference for the location in which this stimulation occurred, using a pharmacogenetic strategy. Juvenile (1-month-old) *Mrgprb4-cre* male mice were injected neonatally with an AAV encoding the hM3-(G_q-coupled) DREADD[19], the activation of which by clozapine-N-oxide (CNO) causes membrane depolarization (Fig. 4a). Calcium imaging experiments confirmed that CNO was able to induce calcium transients in MRGPRB4$^+$ spinal afferent fibres co-expressing GCaMP3.0 and hM3DREADD (Supplementary Fig. 10 and Supplementary Note 7).

We tested whether activation of MRGPRB4$^+$ neurons would promote a preference for the chamber associated with CNO treatment. Because most mice during a pre-training exposure to the CPP apparatus showed an initial preference for one of the two side chambers (Fig. 4b, 'I.P.' and Supplementary Fig. 11), we used a biased design[20,21] to test whether activation of MRGPRB4$^+$ neurons would increase the animals' preference for the initially non-preferred (I.N.P.) chamber. To do this, mice were conditioned over 4 days (experimenter blind to genotype) by pairing a 1-h exposure to CNO with the I.N.P. chamber on each of two days, alternating with exposure to saline in the I.P. chamber (Fig. 4c, lower).

When tested on the day after conditioning, *Mrgprb4-hM3DREADD* mice (Fig. 4e), but not a series of control mice (Fig. 4f–i), exhibited a statistically significant increase in the time they spent in the I.N.P. chamber that was paired with CNO exposure (Fig. 4c, d; 190 ± 95% increase, $P < 0.01$ pre- versus post-training, $n = 15$ mice; see Supplementary Fig. 12 for scatterplots; Supplementary Fig. 13f and Supplementary Note 8). The mean difference score of the experimental animals in the I.N.P. chamber (Fig. 4e, j; 253 ± 65.8) was significantly higher than that of the pooled controls (Fig. 4j; 51.8 ± 35.3, $t = 2.92$, $P < 0.01$, $n = 33$ mice) and our statistical power (0.83) was sufficient to detect this difference given the effect size (difference of means = 201 ± 70; 95% confidence interval: 62.9–339.3; Cohen's $d = 0.872$). There was no statistically significant difference between groups in the time spent in the I.N.P. chamber during the pre-test (Supplementary Fig. 12). These data suggest that artificial activation of MRGPRB4$^+$ neurons *in vivo* is positively reinforcing and/or anxiolytic[18,21]. In contrast, artificial activation of MRGPRD$^+$ neurons using CNO and DREADD produced neither CPP (Fig. 4d, h and Supplementary Fig. 12d and 13d) nor (in separate experiments) conditioned place aversion[22,23] (CPA; data not shown). The failure to obtain CPA using MRGPRD::DREADD mice may reflect technical or biological factors. For example, although hindpaw injection of formalin produces CPA[22], MRGPRD$^+$ neurons are not required for nocifensive responses to formalin[24].

Here we report the first application of calcium imaging to record physiological responses of primary sensory neurons to cutaneous stimulation

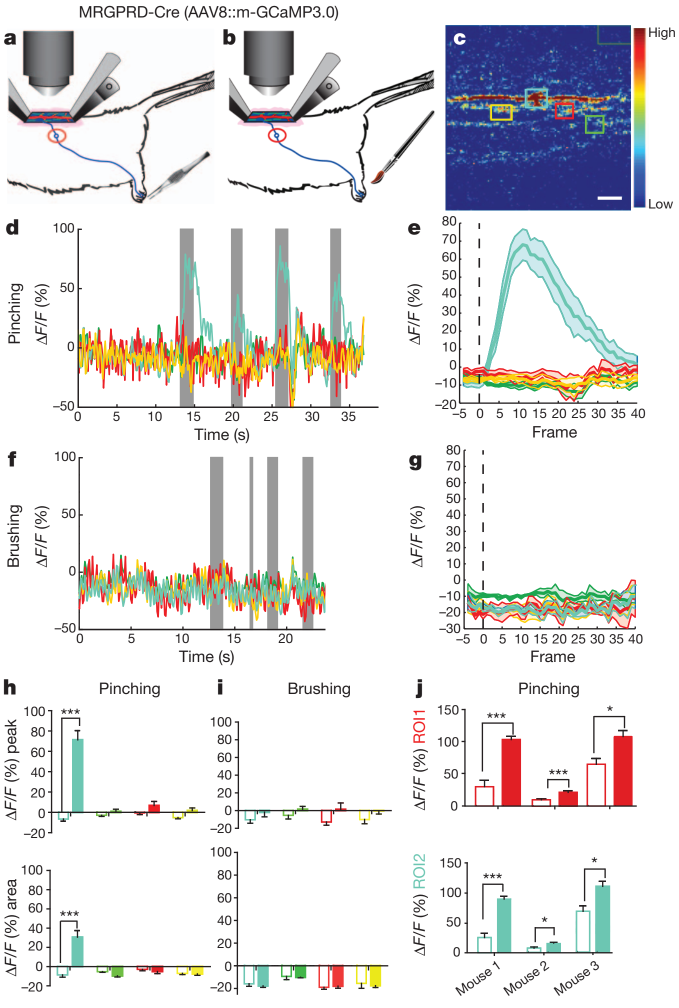


Figure 2 | Activation of MRGPRD fibres by pinching. **a**, **b**, Schematics illustrating pinching (**a**) and stroking (**b**) stimuli. **c**, GCaMP3.0 fluorescence in one imaging frame during stimulation and ROIs used for imaging in **d**–**i**. The green rectangle (upper right) is region for background subtraction. Scale bar, 9 µm. **d**, Superimposed traces from different colour-coded ROIs (**c**) in a single trial consisting of four pinch stimuli (grey bars). **e**, Average response to pinching in a single mouse ($n = 4$ trials, 7 stimuli total). See also Supplementary Fig. 5a–e. **f**, Response to four brushing stimuli (grey bars) delivered to pinch-sensitive digit (**d**), in same ROI (**c**). See also Supplementary Fig. 9g–l. **g**, Average response to brushing ($n = 2$ trials, 7 stimuli total). **h**, **i**, MPI $\Delta F/F_{peak}$ (upper) or integrated area (lower) from curves in **e**, **g**, respectively. Open and filled bars are 5 frames before and 40 frames after stimulus delivery, respectively (see Supplementary Note 9). **j**, MPI $\Delta F/F_{peak}$ in two different ROIs (red and turquoise bars) from each of three mice. Open and filled bars as in panels **h**, **i**. See also Supplementary Table 2. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. All data are mean ± s.e.m.

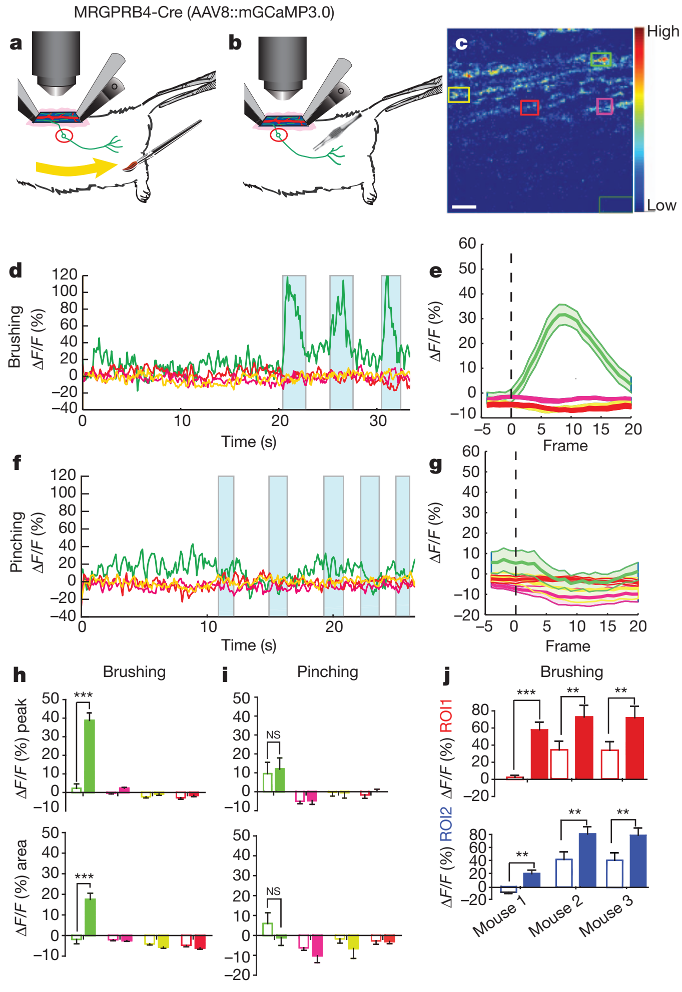


Figure 3 | Activation of MRGPRB4 fibres by stroking. **a**, **b**, Schematics illustrating brushing (**a**) and pinching (**b**) stimuli. **c**, GCaMP3.0 fluorescence in one imaging frame during stimulation and ROIs used for imaging in **d**–**i**. The dark-green rectangle (lower right) is the region used for background subtraction. Scale bar, 8.5 µm. **d**, Superimposed traces from different colour-coded ROIs (**c**) in a single trial of three brush stimuli (turquoise bars). **e**, Average response to brushing from a single mouse ($n = 5$ trials, ~3–6 stimuli per trial). See also Supplementary Fig. 5g–j. **f**, Response to five pinching stimuli (turquoise bars) in brush-sensitive region (**d**), in same ROI (**c**). See also Supplementary Fig. 9a–f. **g**, Average response to pinching from the same animal ($n = 2$ trials, 10 stimuli total). **h**, **i**, MPI $\Delta F/F_{peak}$ (upper) or integrated area (lower) calculated from the curves in **e**, **g**, respectively. Open and filled bars are 5 frames before and 20 frames after stimulus delivery, respectively. NS, not significant. **j**, MPI $\Delta F/F_{peak}$ in two different ROIs (red and blue graphs) from each of three independent mice. Open and filled bars as in panels **h**, **i**. See also Supplementary Table 3. $**P < 0.01$; $***P < 0.001$. All data are mean ± s.e.m.

in an intact animal. Using genetically encoded calcium sensors, we identify a molecularly defined subpopulation of unmyelinated fibres that responds to innocuous stroking of hairy skin *in vivo*. Selective manipulation of these neurons *in vivo* also provides the first example of a genetically identified population of C fibres whose functional activation has a positive[18] rather than negative[23] behavioural valence.

Our results provide proof-of-principle for a means to link molecular identity to stimulus selectivity for primary sensory neuron subtypes that cannot be functionally characterized using more conventional approaches. The inability to detect activation of MRGPRB4 neurons by mechanical stimuli in isolated skin-nerve preparations[10], for whatever reason(s) (see Supplementary Discussion 1), clearly distinguishes them functionally from other populations of unmyelinated mechanosensitive neurons that have been recently characterized in this manner[9].

In humans, C-LTMRs, also called C-tactile (CT) afferents[2], have been associated with gentle, pleasant stroking of hairy skin[25]. Although the unmyelinated axons of MRGPRB4$^+$ neurons[7] indicate that they are C fibres, it is not possible to classify them as low-threshold mechanoreceptors according to the electrophysiological criteria established for C-LTMRs[26], as they are not activated by von Frey filaments. Nevertheless, our imaging and behavioural data taken together suggest that MRGPRB4$^+$ neurons detect massage-like stroking of hairy skin that is pleasant or rewarding, supporting the hypothesis that they may constitute at least one class of CT afferents[7]. Further studies will be required to determine whether sensory neurons with similar properties exist in humans. In mice, activation of these neurons may normally occur during social (affiliative or maternal) interactions, during self-stimulation or other behavioural conditions. Identification of these conditions will enable assessment of the requirement of

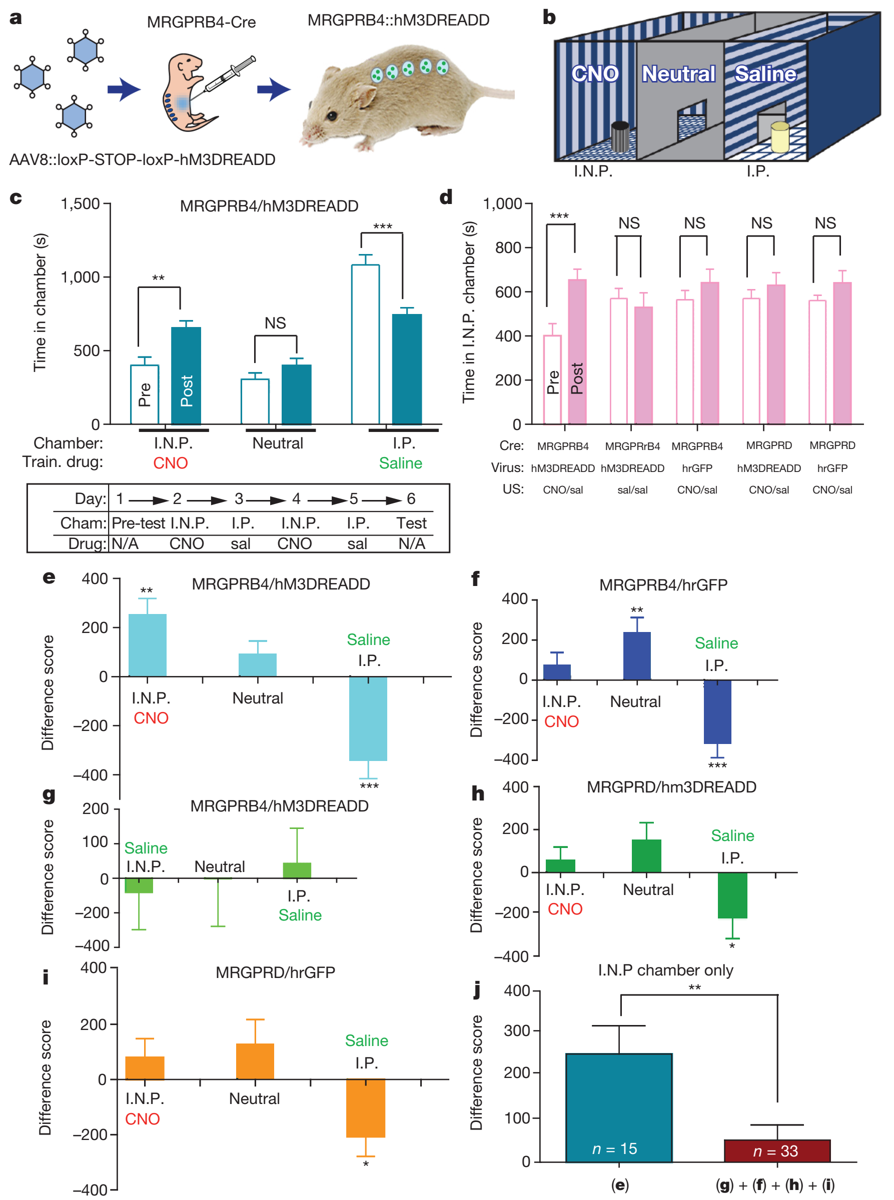


Figure 4 | Activation of MRGPRB4 neurons promotes conditioned place preference. **a**, **b**, Schematic of experiment (**a**) and CPP apparatus (**b**). I.N.P. and I.P. indicate initially non-preferred and preferred chambers, respectively (**c**, schematic, 'pre-test'). **c**, Top: absolute time (s) in each chamber before (open bars; 'pre') versus after (filled bars; 'post') conditioning for the experimental group. Train. drug indicates CNO or saline paired with the indicated chamber. Bottom: schematic of experimental design. Cham., chamber. **d**, Time in I.N.P. chamber for experimental (replotted from panel **c** for direct comparison) and control groups. $^{**}P < 0.01$; $^{***}P < 0.001$; NS, not significant. See Supplementary Note 10 for F values and Supplementary Fig. 13. **e**–**i**, Difference scores ((time in indicated chamber after training) – (time in chamber before training)) for experimental (**e**, $n = 15$) and control (**f**, **g**, **h**, **i**, $n = 9$, 6, 8, 10, respectively) groups. **j**, Comparison of mean difference scores for the I.N.P. chamber for the experimental (**e**) and the pooled control (**f**, **g**, **h**, **i**) groups. There was no significant difference between control groups. $^{*}P < 0.05$, $^{**}P < 0.01$; $^{***}P < 0.001$. All data are mean ± s.e.m.

MRGPRB4^{+} neurons for specific behavioural responses, via loss-of-function experiments. The functional characterization of a novel population of CT afferents in the mouse now opens the way to identifying the molecular transduction mechanisms operating in these cells, and the higher-order circuitry that these neurons engage to produce a positive affective state[27].

METHODS SUMMARY

Mice expressing reporters and/or Cre recombinase targeted to the *Mrgprb4* locus[7] were generated by homologous recombination in embryonic stem cells, according to standard procedures. Heterozygous neonates from MRGPRD-Cre and MRGPRB4-Cre mice were injected intraperitoneally[28] with Cre-dependent AAV8 viruses expressing GCaMP3.0 or mGCaMP3.0 (ref. 14) or hM3-(G_q-coupled) DREADD[19], and imaged as adults (≥8 weeks old).

Electrophysiology experiments on *ex vivo* skin-nerve preparations from adult heterozygous *Mrgprb4-EGFPf* reporter mice were performed as described[26]. For calcium imaging, after a dorsal laminectomy the spinal column was stabilized[29,30] and filled with imaging solution (see Methods). Imaging was performed using a two-photon laser-scanning microscope (Ultima, Prairie Inc.) using an Olympus ×40 0.8 N.A. water immersion objective, at 128 × 128 pixel resolution with an acquisition rate of 8–12 frames per second. Mechanical stimuli were delivered using a custom-modified No. 5 sable paint brush or serrated forceps, in a manner electronically time-stamped to image acquisition (Supplementary Fig. 14 and Methods). Stimuli were delivered to each mouse in a series of trials, separated by a few minutes; each trial consisted of one or more stimuli delivered typically at intervals of several seconds. Chemical stimuli were delivered to the spinal cord using a Pipetman pipette and to the periphery using a syringe pump. Calcium responses were analysed using custom software written in Matlab (see Methods). For calculating $\Delta F/F$ [$(F_{av} - F_0)/F_0$], F_0 is the average of the first ten frames of the recording period.

For behavioural experiments, juvenile (1-month-old) mice neonatally injected with Cre-dependent AAV8 viruses expressing hM3DREADD were subjected to a conditioned place preference assay (CPP[13]) using a biased design[20,21], by an investigator blind to genotype. All mice were tested for their initial chamber preference before conditioning (see Fig. 4c, lower panel).

All data were analysed for statistical significance using repeated measure ANOVAs (unless stated otherwise). After detection of a significant interaction and/or main effect, Bonferroni-corrected post-hoc comparisons of means were performed. Details of statistical analysis are available in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 17 June; accepted 20 November 2012.

1. Morrison, I., Loken, L. S. & Olausson, H. The skin as a social organ. *Exp. Brain Res.* **204,** 305–314 (2010).

2. Olausson, H., Wessberg, J., Morrison, I., McGlone, F. & Vallbo, A. The neurophysiology of unmyelinated tactile afferents. *Neurosci. Biobehav. Rev.* **34,** 185–191 (2010).
3. Dunbar, R. I. The social role of touch in humans and primates: behavioural function and neurobiological mechanisms. *Neurosci. Biobehav. Rev.* **34,** 260–268 (2010).
4. McGlone, F., Vallbo, A. B., Olausson, H., Loken, L. & Wessberg, J. Discriminative touch and emotional touch. *Can. J. Exp. Psychol.* **61,** 173–183 (2007).
5. Dong, X., Han, S., Zylka, M. J., Simon, M. I. & Anderson, D. J. A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. *Cell* **106,** 619–632 (2001).
6. Zylka, M. J., Dong, X., Southwell, A. L. & Anderson, D. J. Atypical expansion in mice of the sensory neuron-specific Mrg G protein-coupled receptor family. *Proc. Natl Acad. Sci. USA* **100,** 10043–10048 (2003).
7. Liu, Q. *et al.* Molecular genetic visualization of a rare subset of unmyelinated sensory neurons that may detect gentle touch. *Nature Neurosci.* **10,** 946–948 (2007).
8. Wessberg, J., Olausson, H., Fernström, K. W. & Vallbo, A. B. Receptive field properties of unmyelinated tactile afferents in the human skin. *J. Neurophysiol.* **89,** 1567–1575 (2003).
9. Li, L. *et al.* The functional organization of cutaneous low-threshold mechanosensory neurons. *Cell* **147,** 1615–1627 (2011).
10. Rau, K. K. *et al.* Mrgprd enhances excitability in specific populations of cutaneous murine polymodal nociceptors. *J. Neurosci.* **29,** 8612–8619 (2009).
11. Zylka, M. J., Rice, F. L. & Anderson, D. J. Topographically distinct epidermal nociceptive circuits revealed by axonal tracers targeted to *Mrgprd*. *Neuron* **45,** 17–25 (2005).
12. Cavanaugh, D. J. *et al.* Distinct subsets of unmyelinated primary sensory fibers mediate behavioral responses to noxious thermal and mechanical stimuli. *Proc. Natl Acad. Sci. USA* **106,** 9075–9080 (2009).
13. Tzschentke, T. M. Measuring reward with the conditioned place preference (CPP) paradigm: update of the last decade. *Addict. Biol.* **12,** 227–462 (2007).
14. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* **6,** 875–881 (2009).
15. Dussor, G., Zylka, M. J., Anderson, D. J. & McCleskey, E. W. Cutaneous sensory neurons expressing the Mrgprd receptor sense extracellular ATP and are putative nociceptors. *J. Neurophysiol.* **99,** 1581–1589 (2008).
16. Edin, B. B., Essick, G. K., Trulsson, M. & Olsson, K. A. Receptor encoding of moving tactile stimuli in humans. I. Temporal pattern of discharge of individual low-threshold mechanoreceptors. *J. Neurosci.* **15,** 830–847 (1995).
17. Panksepp, J. B. & Lahvis, G. P. Social reward among juvenile mice. *Genes Brain Behav.* **6,** 661–671 (2007).
18. Panksepp, J. Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals. *PLoS One* **6,** e21236 (2011).
19. Alexander, G. M. *et al.* Remote control of neuronal activity in transgenic mice expressing evolved G protein-coupled receptors. *Neuron* **63,** 27–39 (2009).
20. Le Foll, B. & Goldberg, S. R. Nicotine induces conditioned place preferences over a large range of doses in rats. *Psychopharmacology (Berl.)* **178,** 481–492 (2005).
21. Cunningham, C. L., Ferree, N. K. & Howard, M. A. Apparatus bias and place conditioning with ethanol in mice. *Psychopharmacology (Berl.)* **170,** 409–422 (2003).
22. Johansen, J. P., Fields, H. L. & Manning, B. H. The affective component of pain in rodents: direct evidence for a contribution of the anterior cingulate cortex. *Proc. Natl Acad. Sci. USA* **98,** 8077–8082 (2001).
23. LaBuda, C. J. & Fuchs, P. N. A behavioral test paradigm to measure the aversive quality of inflammatory and neuropathic pain in rats. *Exp. Neurol.* **163,** 490–494 (2000).
24. Shields, S. D., Cavanaugh, D. J., Lee, H., Anderson, D. J. & Basbaum, A. I. Pain behavior in the formalin test persists after ablation of the great majority of C-fiber nociceptors. *Pain* **151,** 422–429 (2010).
25. Löken, L. S., Wessberg, J., Morrison, I., McGlone, F. & Olausson, H. Coding of pleasant touch by unmyelinated afferents in humans. *Nature Neurosci.* **12,** 547–548 (2009).
26. Woodbury, C. J., Ritter, A. M. & Koerber, H. R. Central anatomy of individual rapidly adapting low-threshold mechanoreceptors innervating the ''hairy'' skin of newborn mice: early maturation of hair follicle afferents. *J. Comp. Neurol.* **436,** 304–323 (2001).
27. Olausson, H. *et al.* Unmyelinated tactile afferents signal touch and project to insular cortex. *Nature Neurosci.* **5,** 900–904 (2002).
28. Foust, K. D., Poirier, A., Pacak, C. A., Mandel, R. J. & Flotte, T. R. Neonatal intraperitoneal or intravenous injections of recombinant adeno-associated virus type 8 transduce dorsal root ganglia and lower motor neurons. *Hum. Gene Ther.* **19,** 61–70 (2008).
29. Davalos, D. *et al.* Stable *in vivo* imaging of densely populated glia, axons and blood vessels in the mouse spinal cord using two-photon microscopy. *J. Neurosci. Methods* **169,** 1–7 (2008).
30. Johannssen, H. C. & Helmchen, F. *In vivo* Ca^{2+} imaging of dorsal horn neuronal populations in mouse spinal cord. *J. Physiol. (Lond.)* **588,** 3397–3402 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Robertson for programming and imaging data analysis; M. Walsh and T. Heitzman for the stimulus delivery system and associated electronics; M. Zelikowsky for help with statistical analysis of behavioural data; H. Inagaki for experimental advice and comments on the manuscript; S. Pease for help with generation of knock-in mice; N. Verduzco, K. Lee and R. Sauza for mice colony maintenance; M. Visel and J. Flannery for training in AAV8 preparation; A. Anderson and C. Pagan for initial experiments using their stereotaxic apparatus; J. Zhang and A. Basbaum for teaching the dorsal laminectomy and for comments on the manuscript; D. Davalos, K. Akassoglou and H. Johanssen for help with the *in vivo* imaging preparation; L. Lagnado and B. Odermatt for SypHy and SyGCamp2 plasmids; C. Shea and M. Martinez for technical assistance; H. Oates-Barker for laboratory management and G. Mancuso for administrative assistance. This work was supported by NIH grants 5PO1NS-48499 and 5R01 NS023476, and by fellowships from EMBO and the Human Frontiers Science Program (S.V.) and the Helen Hay Whitney Foundation (A.M.W.). D.J.A. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions S.V. carried out all imaging and behavioural experiments; A.M.W. helped to configure the two-photon imaging system and developed the light grid method; K.K.R. and H.R.K. carried out electrophysiological recordings in isolated skin-nerve preparations; D.J.A. participated in experimental design and data interpretation and wrote the manuscript together with S.V.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J.A. (wuwei@caltech.edu).

METHODS

Animals. Animals were group-housed, unless otherwise mentioned, at 23 °C with ad libitum access to food and water in a 1-h/11-h light/dark cycle, with the day starting at 07:00. All animal procedures were performed under protocols approved by the Caltech Institutional Animal Care and Use Committee (IACUC).

Generation of *Mrgprb4* knock-in mice. *Mrgprb4-mtdTomato-2A-NLScre-frt-PGK-neo-frt* and *Mrgprb4-EGFPf-2A-FLP-ACN* mice were generated via standard gene-targeting methods in embryonic stem cells, using the previously described 129/SvJ targeting arms of *Mrgprb4* (ref. 7). The lengths of 5′ and 3′ arms were 4.3- and 3.0-kb, respectively. In one construct, the entire open reading frame of *Mrgprb4* (encoded by a single exon) was replaced with an mtdTomato-2A-NLSCre targeting cassette. This cassette was generated as a single open reading frame using overlapping PCR that connected the membrane-tagged tdTomato (containing the 8 amino acids of the MARCKS sequence (MGCCFSKT) fused to the amino terminus of the full-length tdTomato, including its N-terminal methionine[31]) to a nuclear localization signal (NLS)-tagged Cre recombinase via an intervening F2A sequence[32]. This cassette was ligated as a SacII–SalI fragment to the frt-PGK-neo-frt cassette[33]. It was then ligated in-frame to an AscI site at the endogenous ATG start codon of the *Mrgprb4* coding sequence. To generate *Mrgprb4-EGFPf-2A-FLP-ACN* mice, the open reading frame of *Mrgprb4* was replaced by the *EGFPf-2A-FLP* cassette, where EGFPf (farnesylated EGFP; Clontech) was fused via the *2FA* sequence to FLPo (codon optimized FLP recombinase[34]). This cassette was ligated to the self-excising *loxP*-flanked pol II promoter-neomycin resistance cassette (ACN[35]).

Homologous recombination was performed in mouse CJ7 embryonic stem (ES) cells following standard procedures. Correctly targeted ES clones were identified by PCR genotyping of genomic DNA isolated from G418-resistant clones using primer sets flanking the 5′ and 3′ arms of the targeting construct and were further confirmed by Southern blot hybridization using probes that flanked the 5′ and 3′ arms of the targeting construct, as well as an internal probe to exclude illegitimate recombination events. Chimaeric *Mrgprb4-mtdTomato-2A-NLScre-frt-PGK-neo-frt* and *Mrgprb4-EGFPf-2A-FLP-ACN* mice were produced by blastocyst injection of positive ES cells, and heterozygous progeny were generated by mating the chimaeric mice to C57BL/6N mice. Back-crossing to C57BL/6N mice was done for five or more generations.

Neonatal mouse viral injections. P1–P2 pups were removed from their cage and briefly submerged in an ice water bath inside a latex glove with their head up, until they appeared to be anaesthetized (3–5 min). The adequacy of anaesthesia was determined by toe pinch. Pups were then held gently by the head, with padding, the skin of the lower abdomen cleaned with an alcohol swab, and the animals were then immobilized in a plastic gel pocket with their ventral side up. A syringe (insulin syringe, 0.3 cm^3, 8 mm length, 31G needle) was used to inject AAV8 virus (20–25 μl containing 10^{10} AAV8 particles), titred by dot-blot hybridization or by genome copy number (using quantitative real time PCR, qPCR) intraperitoneally (i.p.), avoiding any visible milk spot. The pups were then covered with nesting material and placed on a water circulating heating pad until they began moving. After this recovery period they were returned to their dam and observed for the appearance of a milk spot, indicating that they were healthy and suckling.

Virus production. AAV8 virus particles were produced using crude iodixanol purification as described[36] and concentrated using a Millipore Ultra-15 unit (no. UFC910024).

Immunofluorescence. Adult mice (8–16 weeks old) were anaesthetized with ketamine/xylazine and perfused with 20 ml 0.1 M phosphate buffer solution (PBS; pH 7.4; 4 °C) followed by 25 ml 4% paraformaldehyde (PFA) in PBS (4 °C). Dorsal root ganglia (DRG) were dissected from the perfused mice, post-fixed in 4% PFA at 4 °C for 5 min, cryoprotected in 20% sucrose in PBS at 4 °C for 24 h, and frozen in OCT at −80 °C. Tissues were sectioned at 20 μm with a cryostat. The sections collected on slides were dried at 37 °C for 15 min. The slides were washed with PBS containing 0.2% Triton X-100 (PBT) and blocked with 10% goat/donkey serum in PBT for 30 min. All sections were incubated overnight with primary antibodies diluted in blocking solution at 4 °C. The primary antibodies used were: rabbit anti-GFP (A-11122; Molecular Probes; 1:1,000), rabbit anti-hrGFP (240142; Stratagene; 1:200) and chicken anti-GFP (GFP1020; Aves Labs; 1:1,000). After incubation with primary antibody, sections were washed with PBT and incubated with secondary antibodies at room temperature for 2 h. Secondary antibodies were diluted 1:250 in blocking solution and were conjugated to Alexa 488 or Alexa 568 (Molecular Probes). Sections were counterstained with TO-PRO-3 (Molecular Probes), washed with PBT and mounted with Vectashield. Images were obtained using an Olympus Confocal Microscope system.

Electrophysiological recording in *ex vivo* skin-nerve preparations. The *ex vivo* somatosensory system preparation has been described in detail previously[26]. Briefly, adult *Mrgprb4-EGFP-2A-FLP* mice were anaesthetized with a mixture of ketamine (90 mg kg^{-1}) and xylazine (10 mg kg^{-1}), the skin of the dorsal hindpaw and limb was shaved and then the mice were transcardially perfused with chilled and oxygenated artificial cerebral spinal fluid. Surgical dissection was performed to isolate intact the hemisected spinal cord, L2–L3 dorsal roots and DRGs, saphenous nerves and innervated skin from the left or right hindlimbs. The skin was pinned hairy-side up on an elevated platform, keeping the dermal side perfused and the epidermis dry. Bath temperature was maintained at 31 °C. $EGFP^+$ cells were targeted using fluorescent microscopy and DIC optics. Recording electrodes contained 5% neurobiotin (NB) in 1 M potassium acetate. A small amount of <1% lucifer yellow was added to the solution for better visualization of the microelectrode tip under fluorescent illumination. After impalement of a targeted neuron projecting through the saphenous nerve we first searched for its receptive field by stroking the skin using a fine camel-hair brush. Next the skin was searched using a small glass rod. We next applied thermal stimuli by flooding the skin surface with first cold (0 °C) and then hot (52 °C) buffered saline. Finally, in some of the experiments the skin was then treated with a cocktail of inflammatory compounds (10 μM histamine, 10 μM bradykinin, 10 μM serotonin and 10 μM prostaglandin E_2, in 50% DMSO and 50% buffered Krebs solution at pH 6) for 3–5 min to determine if the cells were either chemosensory or whether they could be sensitized to respond to the other stimulus modalities.

Summary of electrophysiology results. We recorded from 25 $EGFP^+$ positive cells from 10 saphenous nerve preparations made from *Mrgprb4-EGFP-2A-FLP* mice. None of these 25 cells could be activated with mechanical stimulation of the skin. Of these, 21 were also thoroughly tested for thermal sensitivity and were found to be unresponsive. Finally, four cells were tested with mechanical thermal and chemical stimuli (inflammatory soup) and all four remained unresponsive.

Calcium imaging. Mice 2 months or older were sedated by i.p. injection of a mix of ketamine (100 mg kg^{-1}), xylazine (15 mg kg^{-1}), acepromazine (2.5 mg kg^{-1} in 0.9% NaCl). During surgery, body temperature was maintained at 37 °C with a heating blanket.

A dorsal laminectomy was performed mostly at spinal level L2–L4 (but occasionally at L1–L3) as described[30] but without removing the dura. The spinal column was stabilized using Narishige STS-A spinal clamps[29]. In addition we used a head-holding adaptor from Kopf (923-B Mouse Gas Anaesthesia Head Holder) that has installed an anaesthesia/gas mask for positioning the mouse head. In this apparatus the gas is applied through a standard hose barb positioned above the nose on the mask. The inlet fills a large gas chamber around the snout, a second hose barb below the mask is provided for vacuuming off excess, expelled gasses. The animals were maintained under continuous anaesthesia for the duration of the imaging experiments with 1–2% isofluorane or with hourly injections of the above ketamine mix. A well was built around the exposed spinal cord using Gelseal (Amersham Biosciences Corp) and Kwik Sil Adhesive (WPI). Warm imaging solution (in mM: 130 NaCl, 3 KCl, 2.5 $CaCl_2$, 0.6 $6H_2O \cdot MgCl_2$, 10 HEPES without Na, 1.2 $NaHCO_3$, 10 glucose, pH 7.45 with NaOH) (37 °C) was repeatedly applied to prevent drying and maintain tissue integrity, and to allow the use of immersion objectives. During imaging the body temperature of the animals was maintained at 37 °C with a heating blanket and an air-therm heater (WPI) placed inside the microscope area.

Imaging experiments were performed under a two-photon laser-scanning microscope (Ultima, Prairie Instruments Inc.). Live images were acquired at 8–12 frames per second, at depths below the pia ranging from 100 to 250 μm, using an Olympus ×40 0.8 N.A. water immersion objective, at 128 × 128 pixel resolution with a laser tuned to 940 nm wavelength, and emission filters 525/50 nm and 595/50 nm for green and red fluorescence, respectively. Laser power was adjusted to be 20–25 mW at the focal plane (maximally 35 mW), depending on the imaging depth and level of expression of GCaMP3.0. Focal planes containing fibres activated by stimulation of a given peripheral area were identified by trial and error. tdTomato fluorescence was used to identify $MRGPRB4^+$ fibres until photobleaching occurred.

Stimulus delivery during imaging experiments. Brushing stimuli were delivered using a sable paint brush No. 5. Pinching stimuli were delivered using serrated forceps (Adson-Graefe tissue forceps, Fine Science Tools, catalogue no. 11030-12). A touch sensor was designed to allow detection of a finger touch to a conductive band (copper) mounted on the paint brush (Methods and Supplementary Fig. 14a, b). The function of the sensor amplifier/brush is to allow the coordinated movement of the brush tip with a light touch of the sensor band to produce a TTL (+5 V DC) compatible voltage pulse that can be time-stamped to the image acquisition. For the pinching stimulation the touch sensor was modified to detect closure of the forceps. This adjustment was accomplished by mounting a plastic screw on the forceps so that the contact closure occurred at a consistent position (Methods and Supplementary Fig. 14b).

The circuitry inside the touch-sensor box was designed as follows (see circuit diagram in Supplementary Fig. 14c): the stimulus device (brush or forceps) was attached by a small wire with a male pin at the end. A 2-m cable with a matching

female receptacle and a BNC connector conveyed the electrical signal to the touch-sensor amplifier. The probe input on the amplifier was connected to +2.5-V DC through a 10 MΩ resistor. This point was attached to a high impedance follower. The shield (outer part of the coaxial cable) on the probe wire is 'driven' by being connected to the output of the follower. This provides a low-impedance shield to keep electrical interference from coupling to the touch probe input line. When a touch is made, the output of the follower amplifier has a noise envelope (primarily 60/120 Hz) picked up by the body of the person touching the probe band (or the metal body of the forceps). The signal from the follower amplifier is rectified and injected into the positive input of a voltage comparator. The minus input of this circuit is connected to the wiper of a potentiometer on the front panel that provides a sensitivity adjustment. This adjustment allows for the wide range of touch sensitivity that is needed. When the voltage on the plus input (signal from the probe amplifier) exceeds the voltage on the minus input (set by the potentiometer) the output of the comparator is increased. The output of this comparator is conveyed to a BNC connector on the panel as a TTL pulse. The voltage level on this BNC remains high (+5 V) as long as the 'touch' is being made. The signal is internally directed to a three position switch that allows for an LED to be lit or a tone to be generated, enabling visual or auditory confirmation of times when stimulation is performed. The TTL pulse is recorded by the Trigger Sync program (Prairie) which is time-locked with the two-photon image acquisition system (Prairie View, Prairie), thereby identifying imaging frames at which the mechanical stimuli were applied.

We concurrently recorded where on the animal the mechanical stimulation was applied. First, a dim red grid was projected onto the mouse (so as to have the least interference with the detection of the green fluorescence) using a laser pico projector (MicroVision, SHOWWX) to deliminate a coordinate system for stimulation. Then the movement of the brush and the location in the peripheral areas of the brushes were recorded using a camera (Basler, A601f-2).

For delivery of chemical stimuli to the spinal cord, KCl, final concentration (60 mM[37]), α,β-methylATP (5 mM[38]) and CNO (1.5 mM) were delivered manually to the imaging bath using a pipetman pipette.

For delivery of chemical stimuli to the periphery, α,β-methylATP (10 μl from 1 mM solution) and capsaicin (10 μl from 1 mM solution (10% DMSO in saline)) was injected in the ventral and dorsal hindpaw of MRGPRD and MRGPRB4 mice, respectively, using a syringe pump (WPI, Inc., sp200i syringe pump). The timing of the injection was controlled by the two-photon image acquisition system and associated software (Prairie View and Trigger Sync, Prairie Technologies) to link it with image acquisition.

Analysis of imaging data. GCaMP3.0 responses were quantified using custom software written in Matlab (VivoViewer software). Initially, the raw data were filtered by smoothing using a Gaussian filter. The filter is represented by a 3×3 matrix with values proportional to a two-dimensional Gaussian with its peak at the centre, s.d. = 0.5, and normalized so that the matrix's entries sum to 1. The filtered value for each pixel = (its original value × the filter value in the centre) + (original values of the adjacent pixels each multiplied by their corresponding filter values). To calculate values for pixels at the edge of the image, the image is treated as though there are pixels beyond the edge with values equal to those of the nearest edge pixel. Next, the images were subjected to background subtraction to remove excess background noise. This was accomplished by drawing an ROI around a region without any visible structures and calculating the average pixel value in that background ROI, for each frame used for analysis. This value was then subtracted from every pixel in the corresponding frame.

The average fluorescence intensity, F_{av}, was measured by calculating the average (background-subtracted) pixel values in a given ROI, for each image frame recorded during a time interval spanning before and during the stimulation period. The F_{av} was then converted to $\Delta F/F$ using the formula $\Delta F/F = (F_{av} - F_0)/F_0$, where F_0 is the baseline fluorescence value, measured as the average pixel intensity during the first 2–11 frames of each imaging experiment. The resulting time series of $\Delta F/F$ in a given ROI was smoothed using a moving average with a window of three frames. For a window of size M the following equation is used: for a time series, f, of N frames and a window size of M for the moving average (where M is an odd integer), the nth term of the new time series, F, is given by $F_n = \sum_{i=n-m}^{n+m} \frac{f_i}{2m+1}$ where $m = \frac{M-1}{2}$ if both $n > \frac{M-1}{2}$ and $n \leq N - \frac{M-1}{2}$ are true. Otherwise, $m = \min\{n-1, N-n\}$.

For calculation of the trial average curves (see, for example, Figs 2e, g and 3e, g) for mechanical stimuli we used a seven-frame smoothing window. Sections of the $\Delta F/F$ time series during which a stimulus occurred were collected for multiple trials, aligned to the onset of the stimulus, and averaged to find the mean response curve. Because repeated mechanical stimuli were delivered during each experimental trial, to be consistent each $\Delta F/F$ trace was calculated for a period of five frames just before each stimulus onset and for the subsequent 20 or 40 frames (that is, the first frame of these 20–40 frame series coincided with the initiation of the stimulus). From these values we calculated the mean peak $\Delta F/F$ (MPI $\Delta F/F_{peak}$) and area under the curve for all the applied stimuli across trials. The average $\Delta F/F$ values for specific ROIs in the same field of view were tested for statistical significance by repeated measures ANOVA, followed by Bonferroni-corrected post-hoc comparison of means.

In the case of chemical stimulation (delivered either to the spinal cord or to the periphery), typically a single trial was performed for a given mouse, due to the difficulty of maintaining the same focal plane during the period of application of the chemical to the spinal cord or the period required for diffusion of the liquid bolus delivered for peripheral injection, respectively. In these cases, therefore, the MPI $\Delta F/F_{peak}$ before and during the stimulation period were calculated for multiple mice imaged using ROIs of similar size, and were compared for statistical significance (relative to pre-stimulus baseline) by repeated measures ANOVA, followed by Bonferroni-corrected post-hoc comparison of means (unless stated otherwise).

The $\Delta F/F$ values in supplementary Fig. 7b and e were corrected for photobleaching as described[39].

Behaviour. The conditioned place preference (CPP) protocol was based on previous studies[13,17]. As we tested for a positive valence effect of activation of MRGPRB4 neurons, we used a biased compartment assignment procedure in which activation of the neurons is tested for its ability to increase the time spent in the initially non-preferred chamber[20]. The CPP apparatus consisted of a rectangular chamber divided into three compartments ($300 \times 150 \times 150$ mm per compartment), connected via an opening (50×50 mm) in each delimiting wall. The two side (test) compartments were designed to have different visual and tactile cues, by having distinct walls (horizontal or vertical alternating white and black stripes) and distinct floors (different shapes of floor grids with big or small square holes). In addition a 1-inch-diameter polyvinylchloride (PVC) pipe coupler (two schedule 40 wall thickness), either threaded or smooth, was placed in the centre of each side compartment to enrich for tactile cues[17]. The centre compartment was a neutral plastic enclosure (see Fig. 4b). This design was chosen so as to promote a compartment preference assignment for each mouse. A video tracking system (Noldus Ethovision) recorded all animal movements.

Because our hypothesis is based on the social reward mediated by social contact in juvenile mice[17], the mice used were approximately 1-month old. After weaning the mice were maintained in social groups and left undisturbed until the start of the CPP assay. The paradigm was completed in 6 days. The day before pre-testing the mice were socially isolated in their home cage. On day 1 of the procedure each mouse was placed in the central compartment and allowed to explore the entire apparatus freely for 30 min (pre-test). After the pre-test the initial preference of each mouse for a given side compartment was recorded. With our apparatus design most of the mice showed an initial preference for one of the two side compartments. Conditioning was initiated on day 2 and encompassed four sessions performed on four consecutive days. In the first conditioning session mice were injected i.p. with CNO ($5\,\mathrm{mg\,kg^{-1}}$)[19] (or saline of an equivalent volume for some control mice) and placed for 1 h (based on the observation that CNO effects peak between 45 and 50 min after administration[19]) in the initially non-preferred (I.N.P.) compartment. On day 3, during the second conditioning session, all mice were injected with saline and confined for 1 h in the opposite (that is, initially preferred, I.P.) compartment. (The second conditioning session was performed the following day as CNO effects last for 9 h (ref. 19).) On day 4 and day 5 the first and second conditioning sessions were repeated, respectively. The time between the i.p. injections of CNO or saline and the placement of the mice in the compartment was between 5–10 min, which is compatible with the time that is needed for CNO to start having an effect[19]. On day 6, the mice were tested for their side compartment preference by placing them in the centre compartment and allowing them to explore the entire apparatus freely for 30 min (post test). All sessions were conducted blind to the genotype/injected virus of each mouse. For the conditioned place aversion (CPA assay) the mice remained group-housed until the day before the pre-test. After the pre-test, on the first day of the conditioning session the mice were injected with saline and confined in the I.N.P. compartment. On the second day of conditioning the mice were injected with CNO (except for the saline control mice) and placed in the I.P. compartment. On the third and fourth day of conditioning the first and second sessions of conditioning were repeated, respectively. On the sixth day the mice were tested for their preference in the three-compartment arena.

Behavioural data analysis. Difference scores for each chamber (time in chamber during post-test minus time in chamber during pre-test) were analysed for statistical significance (significant difference from zero) using simple or repeated one-way ANOVA ($P < 0.05$) followed by a Bonferroni-corrected post-hoc comparison of means. For the comparison of the mean difference scores between the experimental and the pooled control groups (Fig. 4j), an unpaired t-test was used.

Statistical analysis of all other metrics was performed using a repeated two-way mixed model ANOVA (unless otherwise stated) (with each group as the between-subject variable and pre-training versus post-training as the within-subject variable). Detection of a significant interaction and/or main effect was followed by Bonferroni-corrected post-hoc comparison of means.

Drugs. Clozapine-N-oxide (CNO) was obtained from Biomol International, and dissolved in saline.

31. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45,** 593–605 (2007).
32. de Felipe, P. *et al.* E unum pluribus: multiple proteins from a self-processing polyprotein. *Trends Biotechnol.* **24,** 68–75 (2006).
33. Meyers, E. N., Lewandoski, M. & Martin, G. R. An Fgf8 mutant allelic series generated by Cre- and Flp-mediated recombination. *Nature Genet.* **18,** 136–141 (1998).
34. Raymond, C. S. & Soriano, P. ROSA26Flpo deleter mice promote efficient inversion of conditional gene traps *in vivo*. *Genesis* **48,** 603–606 (2010).
35. Bunting, M., Bernstein, K. E., Greer, J. M., Capecchi, M. R. & Thomas, K. R. Targeting genes for self-excision in the germ line. *Genes Dev.* **13,** 1524–1528 (1999).
36. Zolotukhin, S. *et al.* Recombinant adeno-associated virus purification using novel methods improves infectious titer and yield. *Gene Ther.* **6,** 973–985 (1999).
37. Hamada, F. N. *et al.* An internal thermal sensor controlling temperature preference in *Drosophila*. *Nature* **454,** 217–220 (2008).
38. Hamilton, S. G., McMahon, S. B. & Lewin, G. R. Selective activation of nociceptors by P2X receptor agonists in normal and inflamed rat skin. *J. Physiol. (Lond.)* **534,** 437–445 (2001).
39. Berry, J. A., Cervantes-Sandoval, I., Nicholas, E. P. & Davis, R. L. Dopamine is required for learning and forgetting in *Drosophila*. *Neuron* **74,** 530–542 (2012).

doi:10.1038/nature11729

NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice

Michael T. Heneka[1,2]*, Markus P. Kummer[1], Andrea Stutz[3], Andrea Delekate[4], Stephanie Schwartz[1], Ana Vieira-Saecker[1], Angelika Griep[1], Daisy Axt[1], Anita Remus[4], Te-Chen Tzeng[5], Ellen Gelpi[6], Annett Halle[7], Martin Korte[4,8], Eicke Latz[2,3,5]* & Douglas T. Golenbock[5]*

Alzheimer's disease is the world's most common dementing illness. Deposition of amyloid-β peptide drives cerebral neuroinflammation by activating microglia[1,2]. Indeed, amyloid-β activation of the NLRP3 inflammasome in microglia is fundamental for interleukin-1β maturation and subsequent inflammatory events[3]. However, it remains unknown whether NLRP3 activation contributes to Alzheimer's disease *in vivo*. Here we demonstrate strongly enhanced active caspase-1 expression in human mild cognitive impairment and brains with Alzheimer's disease, suggesting a role for the inflammasome in this neurodegenerative disease. *Nlrp3*$^{-/-}$ or *Casp1*$^{-/-}$ mice carrying mutations associated with familial Alzheimer's disease were largely protected from loss of spatial memory and other sequelae associated with Alzheimer's disease, and demonstrated reduced brain caspase-1 and interleukin-1β activation as well as enhanced amyloid-β clearance. Furthermore, NLRP3 inflammasome deficiency skewed microglial cells to an M2 phenotype and resulted in the decreased deposition of amyloid-β in the APP/PS1 model of Alzheimer's disease. These results show an important role for the NLRP3/caspase-1 axis in the pathogenesis of Alzheimer's disease, and suggest that NLRP3 inflammasome inhibition represents a new therapeutic intervention for the disease.

The chronic deposition of amyloid-β stimulates the persistent activation of microglial cells in Alzheimer's disease[1]. Increased interleukin (IL)-1β amounts have been implicated in the response to amyloid-β deposition[2]. IL-1β is produced as a biologically inactive pro-form and requires caspase-1 for activation and secretion. Caspase-1 activity is controlled by inflammasomes, sensors of microbial components and sterile danger signals. The NLRP3 inflammasome has been implicated in several chronic inflammatory diseases as it can sense inflammatory crystals and aggregated proteins, including amyloid-β[3,4]. Because of the possibility that the neuroinflammatory component of Alzheimer's disease involves inflammasome activation, we assessed the cleavage of caspase-1 in brains from patients with Alzheimer's disease, early-onset Alzheimer's disease and mild cognitive impairment. We observed substantially increased amounts of cleaved caspase-1 in hippocampal or cortical lysates from those patients' brains compared with controls (Fig. 1a and Supplementary Fig. 1), consistent with chronic inflammasome activation[4]. This increase of caspase-1 processing was mirrored in aged APP/PS1 transgenic mice (Fig. 1b). APP/PS1 mice express a human/mouse chimaeric amyloid precursor protein and human presenilin-1, each carrying mutations associated with familial Alzheimer's disease[5], leading to the chronic deposition of amyloid-β, neuroinflammation and cognitive impairment.

Nlrp3$^{-/-}$ mice were crossed into APP/PS1 mice to obtain APP/PS1/*Nlrp3*$^{-/-}$ mice to assess the contribution of the NLRP3 inflammasome to the pathogenesis of Alzheimer's disease. In APP/PS1/*Nlrp3*$^{-/-}$ mice, caspase-1 cleavage was absent. Total brain IL-1β amounts were similar to those in wild-type (WT) animals (Fig. 1b, c). Immunohistochemistry for the inflammasome component ASC detected microglial 'speck' formation in activated (Iba1^{+}) microglia cells from APP/PS1 mice, consistent with inflammasome activation (Fig. 1d). We assessed spatial memory formation in age-matched 16-month-old WT, *Nlrp3*$^{-/-}$, *Casp1*$^{-/-}$, APP/PS1, APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice using the Morris water-maze test including probe trial testing. As expected, aged APP/PS1 mice showed severe deficits in spatial memory formation. However, APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice were largely protected from spatial memory impairment (Fig. 1e, f and Supplementary Figs 2–9). These results were supported by object recognition memory testing (Supplementary Fig. 10). Again, NLRP3- or caspase-1-deficient APP/PS1 mice were protected from memory deficits (Supplementary Fig. 10). To assess the effect of NLRP3 or caspase-1 gene deficiency on neuronal function in murine Alzheimer's disease, we determined hippocampal synaptic plasticity, which is considered to represent the basis of newly formed declarative memories and is often analysed by measuring long-term potentiation (LTP)[6,7]. NLRP3 or caspase-1 deficiency completely prevented LTP suppression in hippocampal slices from APP/PS1 mice (Fig. 1g and Supplementary Fig. 11). Baseline synaptic transmission and short-term plasticity (measured as paired-pulse facilitation) were unaltered (Supplementary Fig. 12). An analysis of spine morphology showed a small but statistically significant reduction of spine density in the pyramidal neurons of APP/PS1 mice, which was prevented by NLRP3 or caspase-1 deficiency (Supplementary Fig. 13). The small degree of spine density reduction suggested that LTP suppression in APP/PS1 mice was primarily mediated by functional rather than structural changes. Body weight and blood glucose amounts were similar between groups of mice (Supplementary Fig. 14). Behavioural analysis in the open field arena verified increased locomotion and slowed habituation in APP/PS1 mice, similar to previous reports[8]. APP/PS1/*Nlrp3*$^{-/-}$ mice, however, had a reduced hyperdynamic phenotype and normalized habituation (Fig. 1h and Supplementary Fig. 15), suggesting that NLRP3 deficiency improved neurobehavioural disturbances such as Alzheimer's-disease-like psychomotor disinhibition. These results support a fundamental role for NLRP3/caspase-1-mediated inflammation in behavioural and cognitive dysfunction in Alzheimer's disease.

Alzheimer's-disease-associated inflammation interferes with APP metabolism and with mechanisms of amyloid-β aggregation and clearance at several levels[9,10]. Thioflavin S staining showed a marked decrease in hippocampal and cortical amyloid-β deposition in the APP/PS1/*Nlrp3*$^{-/-}$ mice (Fig. 2a, b and Supplementary Fig. 16). Additionally, APP/PS1/*Nlrp3*$^{-/-}$ mice showed a 70% reduction in brain concentrations of highly aggregated, formic-acid-extractable

[1]Clinical Neuroscience Unit, Department of Neurology, University of Bonn, Sigmund-Freud-Strasse 25, 53127 Bonn, Germany. [2]Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), 53175 Bonn, Germany. [3]Institute of Innate Immunity, University of Bonn, 53127 Bonn, Germany. [4]Division of Cellular Neurobiology, Zoological Institute, Technische Universität Braunschweig, 38106 Braunschweig, Germany. [5]Department of Medicine and Division of Infectious Diseases and Immunology, University of Massachusetts Medical School, Worcester 01605, Massachusetts, USA. [6]Neurological Tissue Bank, University of Barcelona-Hospital Clinic, IDIBAPS, 08036 Barcelona, Spain. [7]Center for Advanced European Studies and Research-CAESAR, 53175 Bonn, Germany. [8]Helmholtz-Center for Infection Research, HZI, AG NIND, 38124 Braunschweig, Germany.
*These authors contributed equally to this work.

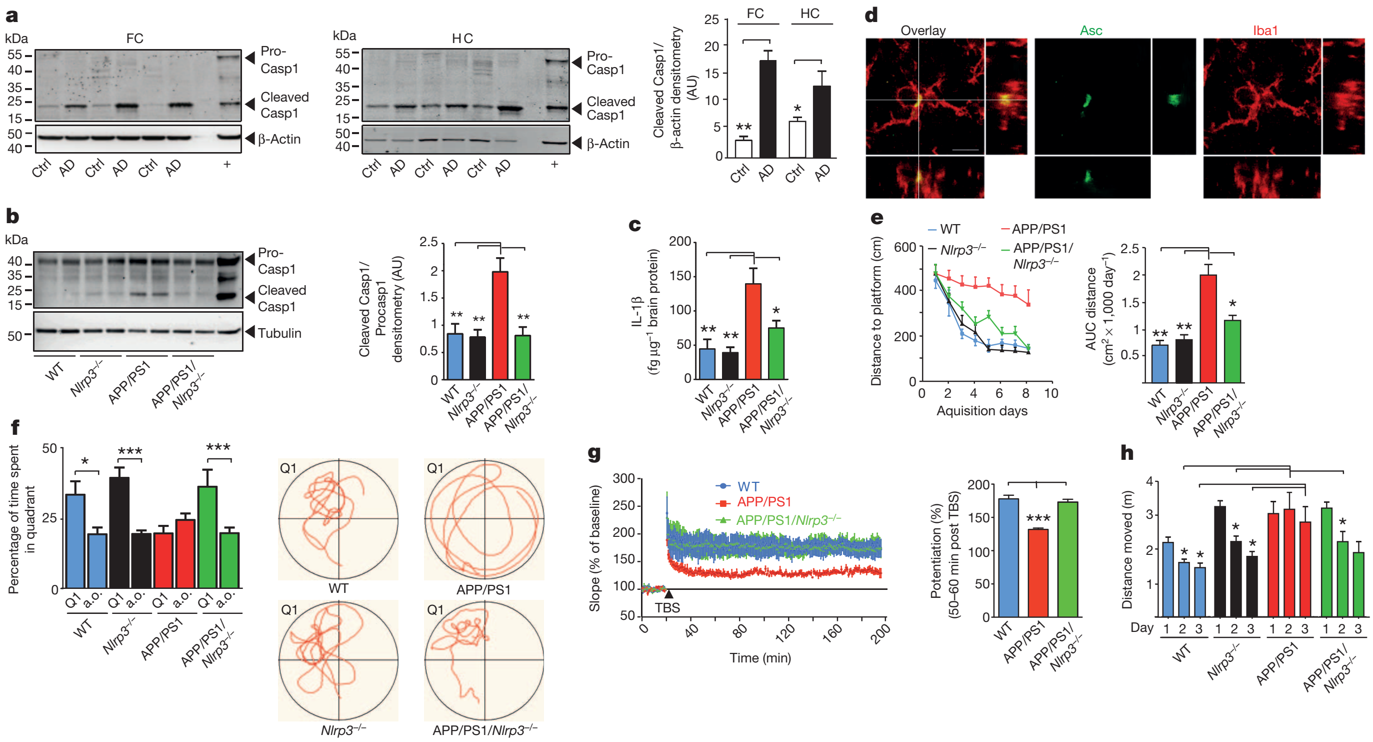


Figure 1 | Protective effects of NLRP3 gene deficiency in APP/PS1 mice on memory and behaviour. **a**, Western blot and quantification of cleaved caspase-1 in brain lysates from frontal cortex (FC) and hippocampus (HC) of patients with Alzheimer's disease (AD, $n = 12$) and controls (Ctrl, $n = 8$) (mean ± s.e.m., Student's *t*-test, $*P < 0.05$, $**P < 0.01$; + is positive control). **b**, Western blot of cleaved caspase-1 and quantification in mice at 16 months ($n = 5$, mean ± s.e.m., analysis of variance (ANOVA), Tukey's (*post hoc*) test, $**P < 0.01$). AU, arbitrary units. **c**, Parenchymal IL-1β in mouse brains from **b** ($n = 5$, mean ± s.e.m., ANOVA, Tukey's test, $*P < 0.05$, $**P < 0.01$). **d**, Immunohistochemistry of microglia from APP/PS1 mice for Iba1 (red) and Asc (green). Scale bar, 10 μm. **e**, Morris water-maze analysis as distance travelled (centimetres) and integrated distance (AUC) for WT ($n = 16$), *Nlrp3*$^{-/-}$ ($n = 12$), APP/PS1 ($n = 14$) and APP/PS1/*Nlrp3*$^{-/-}$ ($n = 15$) mice (mean ± s.e.m., ANOVA, Tukey's test, $*P < 0.05$, $**P < 0.01$). **f**, Probe trial day 9. Q1, quadrant where platform was located on days 1–8. Time spent in all other (a.o.) quadrants was averaged for all of the above mice (mean ± s.e.m.; one-way ANOVA, Tukey's test, $*P < 0.05$, $***P < 0.001$). Representative runs (right panels). **g**, LTP was induced by theta-burst stimulation (TBS) 20 min after baseline recordings in hippocampal slices from mice. LTP is expressed as percentage potentiation of baseline (100%) and the significance is determined 55–60 min after the TBS was given (mean of WT $n = 16$, APP/PS1 $n = 23$, APP/PS1/*Nlrp3*$^{-/-}$ $n = 16$; hippocampal slices measured ± s.e.m. from $n = 6$–9 animals per group; ANOVA, Tukey's test, $***P < 0.001$). **h**, Open field test, age = 16 months. Vertical locomotor activity (distance travelled) decreased over three consecutive days in WT ($n = 16$) and *Nlrp3*$^{-/-}$ ($n = 12$) mice. Habituation was not observed in APP/PS1 mice showing a hyperdynamic behavioural phenotype. APP/PS1/*Nlrp3*$^{-/-}$ mice ($n = 15$) were indistinguishable from *Nlrp3*$^{-/-}$ mice ($n = 12$) (mean ± s.e.m., ANOVA, Tukey's test, $*P < 0.05$).

forms of amyloid-β (Fig. 2c). This reduction was most probably not due to changes of APP expression and processing, as formation of carboxy (C)-terminal fragments or amounts of β-secretase-1 (BACE1) messenger RNA (mRNA) and protein (Fig. 2 and Supplementary Fig. 17) were unaffected in NLRP3 knockout mice. This conclusion was further strengthened by the analysis of 4-month-old APP/PS1 mice; neither NLRP3 nor caspase-1 deficiency influenced amyloid-β amounts (Supplementary Fig. 18). However, aggregated forms of amyloid-β were markedly reduced in the APP/PS1/*Nlrp3*$^{-/-}$ mice, as shown by the further quantification of amyloid-β species by enzyme-linked immunosorbent assay (ELISA). These studies showed a strong reduction of amyloid-β_{1-40} and amyloid-β_{1-42} in the APP/PS1/*Nlrp3*$^{-/-}$ mice after sequential extraction by radio-immunoprecipitation assay (RIPA) and sodium dodecyl sulphate (SDS) buffer, allowing the analysis of soluble and insoluble amyloid-β (Fig. 2d). At 16 months of age, amyloid-β_{1-38} was only detectable in SDS extracts. Again, analysis of the APP/PS1/*Nlrp3*$^{-/-}$ mice showed reduced amounts compared with APP/PS1 mice (Supplementary Fig. 19). Analysis of cerebral amyloid-β amounts of APP/PS1 and APP/PS1/*Casp1*$^{-/-}$ mice showed that caspase-1 deficiency resulted in similar changes in amyloid-β, suggesting that NLRP3 acts through caspase-1 to exert the observed effects (Supplementary Figs 20 and 21).

As both amyloid-β and IL-1β have been implicated in the suppression of LTP[11–13], their reduction may jointly contribute to the protection of LTP, improved spatial memory and normalized behaviour in the NLRP3-deficient APP/PS1 mice.

Microglia are found in increased numbers in close proximity to amyloid-β plaques in Alzheimer's disease. Microglia assembly in the vicinity of plaques is interpreted as an attempt to clear the pathological deposits of amyloid-β through phagocytosis and degradation. The functional impact of phagocytosis is highlighted by studies showing that restricting microglial accumulation and phagocytosis increases amyloid-β deposition[14,15]. As Alzheimer's disease progresses, microglial cells adopt a chronically activated phenotype. Cytokines, including IL-1β, were found to impair microglial clearance functions[16,17]. Likewise, suppression of inflammatory cytokine production resets microglial phagocytosis in APP/PS1 mice[17]. We analysed the impact of NLRP3 or caspase-1 deficiency on the phagocytic capacity of microglia *in vivo* because immunohistochemistry showed microglial ASC and NLRP3 expression (Fig. 1d and Supplementary Figs 22 and 23a). In addition, we observed that inflammasome activation occurred in an age- and amyloid-β deposition-related fashion (Supplementary Figs 23b and 24). We administered a fluorescent derivative of Congo red known as methoxy-XO4, which crosses the blood–brain barrier and has nanomolar binding affinity for amyloid-β. We injected methoxy-XO4 into adult APP/PS1, APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice. Three hours after injection, methoxy-XO4 fluorescence in brain homogenates did not differ between groups (Supplementary Fig. 25). At this

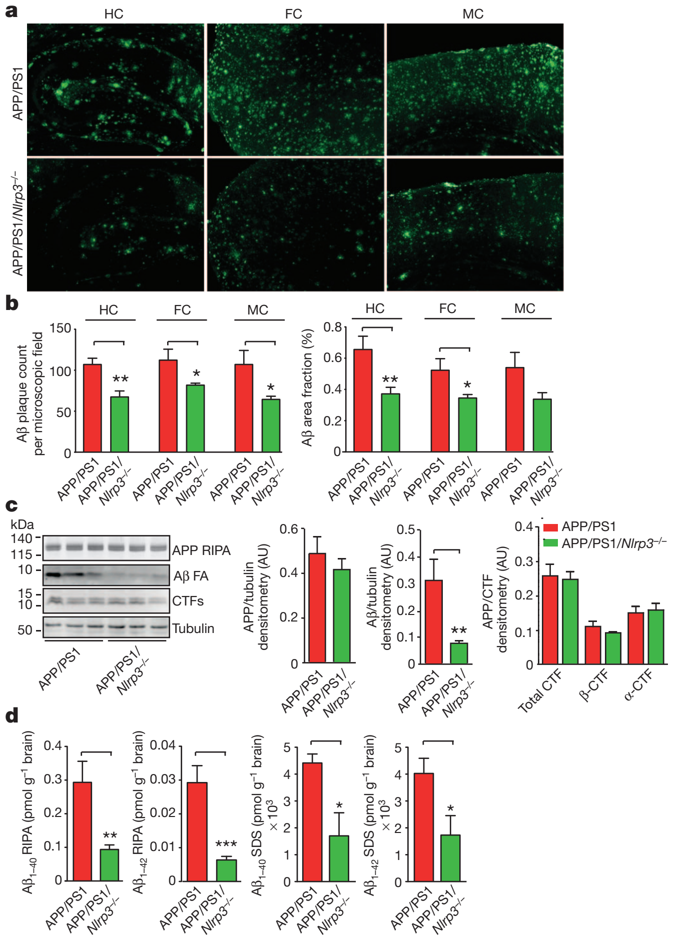


Figure 2 | NLRP3 gene deficiency leads to decreased amyloid-β amounts and deposition. **a**, Amyloid-β plaque deposition was quantified in the hippocampus (HC), frontal cortex (FC) and motor cortex (MC) using thioflavin S. **b**, Quantification of number and surface area of amyloid-β (Aβ) plaques was performed in five consecutive sections per animal and is given as count per area or area fraction (%) ($n = 7$–8, mean ± s.e.m., Students *t*-test, $*P < 0.05$, $**P < 0.001$). **c**, Western blot analysis of RIPA and formic-acid (FA) brain extracts of 16-month-old APP/PS1 ($n = 3$) and APP/PS1/*Nlrp3*$^{-/-}$ ($n = 3$) mice. Densitometrical quantification of APP, formic-acid-soluble amyloid-β and carboxyl-terminal fragments (CTFs) as ratios ($n = 5$, mean ± s.e.m., Student's *t*-test, $**P < 0.01$). **d**, ELISA of RIPA and SDS fractions for amyloid-β_{1-40} and amyloid-β_{1-42} from 16-month-old mice ($n = 5$, mean ± s.e.m., Student's *t*-test, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$).

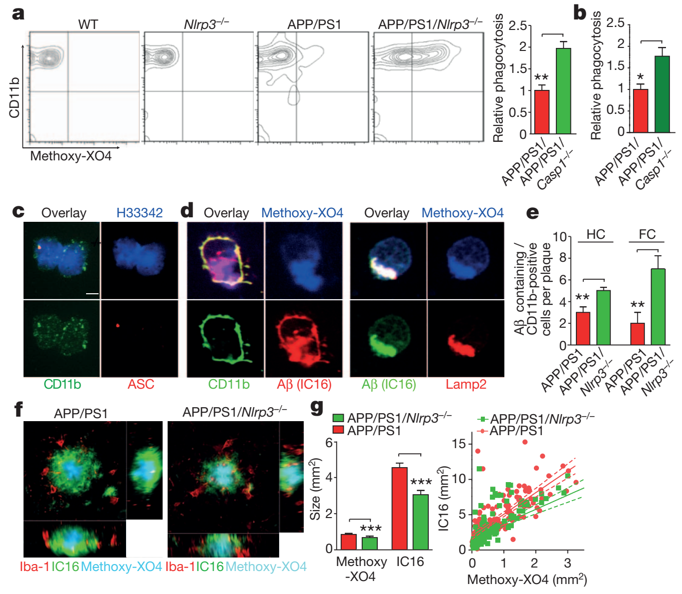


Figure 3 | NLRP3 or caspase-1 deficiency increases microglial amyloid-β phagocytosis. **a**, Quantification of amyloid-β phagocytosis by flow cytometry of microglia isolated from adult mice 3 h after intraperitoneal injection of methoxy-XO4 ($n = 5$, mean ± s.e.m., ANOVA, Tukey's test, $**P < 0.01$). **b**, Same as **a** with APP/PS1 and APP/PS1/*Casp1*$^{-/-}$ mice ($n = 5$, mean ± s.e.m., ANOVA, Tukey's test, $*P < 0.05$). **c**, Immunohistochemistry staining of ASC specks in CD11b^{+} microglia. H33342 is a nuclear stain. **d**, Immunocytochemistry of monoclonal antibody IC16 (anti-amyloid-β), methoxy-XO4 labelled amyloid-β within Lamp2^{+} intracellular structures in CD11b^{+} microglia from 16-month-old APP/PS1 mice. **e**, Quantification of CD11b^{+}, amyloid-β^{+} microglia in the hippocampus (HC) and frontal cortex (FC) of 16-month-old mice ($n = 5$, mean ± s.e.m., Student's *t*-test, $**P < 0.01$). **f**, Representative micrographs from methoxy-XO4-treated APP/PS1 and APP/PS1/*Nlrp3*$^{-/-}$ mice stained for Iba-1 and amyloid-β. **g**, Average IC16-positive amyloid-β plaque size, determined by co-labelling with methoxy-XO4, was markedly reduced in APP/PS1/*Nlrp3*$^{-/-}$ mice (150 plaques were assessed from each group of four mice, mean ± s.e.m., Student's *t*-test, $***P < 0.001$). A scatter plot of all plaques that was analysed by linear regression is shown at the right (150 plaques per group; lines: linear regression analysis; dashed lines: 95% confidence intervals, $R^2 = 0.5588$ for APP/PS1 and $R^2 = 0.4431$ for APP/PS1/*Nlrp3*$^{-/-}$ mice).

point, mice were euthanized and microglial cells were isolated and analysed for methoxy-XO4 fluorescence by flow cytometry. An increase of nearly twofold in amyloid-β phagocytosis was found in APP/PS1/*Nlrp3*$^{-/-}$ or APP/PS1/*Casp1*$^{-/-}$ compared with APP/PS1 mice (Fig. 3a, b), suggesting that NLRP3/caspase-1 inflammasome activation reduces amyloid-β phagocytosis. Microglia were isolated from brains by cytospin. Co-immunostaining again showed microglial ASC speck formation (Fig. 3c). Methoxy-XO4 labelled amyloid-β was detected within CD11b^{+} microglia and co-localized to Lamp2-positive, amyloid-β-containing lysosomes (Fig. 3d and Supplementary Fig. 26). Notably, increased uptake of methoxy-XO4 labelled amyloid-β was associated with enhanced CD36 expression (Supplementary Fig. 27). Although the methoxy-XO4 assay cannot functionally distinguish between increased phagocytosis and impaired degradation, we performed further microscopy. This showed that NLRP3 deficiency substantially altered the characteristics of amyloid-β plaque deposition (Fig. 3e, f and Supplementary Fig. 28). First, the total volume of the amyloid-β plaque was reduced in APP/PS1/*Nlrp3*$^{-/-}$ mice compared with APP/PS1 mice (Fig. 3g). Second, APP/PS1/*Nlrp3*$^{-/-}$ mice showed more reduction in the outer parts of the amyloid-β plaque than in the core. Furthermore, microglial cells surrounding amyloid-β plaques in APP/PS1 mice phagocytosed amyloid-β to a lesser extent (Fig. 3e, g and Supplementary Fig. 29). Together with the documented suppression of microglial phagocytosis by proinflammatory cytokines, these data argue for an increase in phagocytosis in APP/PS1/*Nlrp3*$^{-/-}$ mice. These results may be surprising, because they seemingly contradict a report that experimental local overproduction of IL-1β reduced amyloid-β deposition[18]. However, there are two explanations for these apparently opposite results. First, the NLRP3/caspase-1-axis may use substrates other than IL-1β to constrain microglial amyloid-β phagocytosis. Second, it is likely that the experimental approach that was used by Shaftel *et al.* disrupted the blood–brain barrier, allowing amyloid-β removal by peripherally derived myeloid cells[19]. Similar effects on amyloid-β plaque metabolism have been observed after whole-body irradiation, which also leads to disruption of the blood–brain barrier[1]. Shielding the APP/PS1 brain from radiation restricted the infiltration of peripheral cells to a level that did not significantly contribute to the clearance of parenchymal amyloid-β[20].

In addition to phagocytosis, microglia also contribute to amyloid-β clearance through proteolytic enzymes, including insulin-degrading enzyme (IDE) and neprilysin (NEP)[21]. Cerebral homogenates from APP/PS1/*Nlrp3*$^{-/-}$ or APP/PS1/*Casp1*$^{-/-}$ mice demonstrated an increase of IDE whereas NEP amounts remained unchanged (Fig. 4a and Supplementary Fig. 30). Microglial cells purified from

16-month-old mice were one source of increased IDE transcription (Fig. 4b). Previous work established that a twofold increase of IDE expression is sufficient to reduce amyloid-β deposition strongly[22]. It is likely that the IDE increase enhances the degradation of amyloid-β and the overall amyloid-β reduction in inflammasome-deficient mice. These data suggest that NLRP3 activation negatively affects the microglial clearance function in Alzheimer's disease. Notably, recent evidence suggests that impaired clearance may be the driving force behind sporadic Alzheimer's disease[23], which constitutes the overwhelming majority of cases of human Alzheimer's disease.

Prolonged exposure to amyloid-β leads to persistent activation of microglial cells in Alzheimer's disease. On the basis of gene expression profiles, activated microglial cells may be divided into several different populations. The M1 and M2 subtypes represent the extremes of the range. Markers of alternatively activated microglia of the M2 subtype[24], including 'found in inflammatory zone 1' (FIZZ1) (Supplementary Fig. 31), arginase-1 and interleukin-4, showed increased expression in APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice (Fig. 4c–e). In contrast, cerebral nitric oxide synthase 2 (NOS2), a hallmark of the classically activated M1 proinflammatory phenotype, was reduced in inflammasome-deficient APP/PS1 mice (Fig. 4f, g). Thus, NLRP3- or caspase-1-deficiency results in a skewing of activated microglial cells towards an M2-like activated state. This M2 phenotype is also characterized by increased amyloid-β clearance and enhanced tissue remodelling. In Alzheimer's disease, the upregulation of NOS2 results in tyrosine nitration of several proteins, including amyloid-β, thereby accelerating its aggregation and seeding of new plaques[25]. In agreement with this, APP/PS1/*Nlrp3*$^{-/-}$ mice had less nitrated amyloid-β and a reduced average plaque size as well as fewer nitrated plaque cores (Fig. 4h–j). Because NO and nitrated amyloid-β act as strong LTP suppressors[25,26], a reduction of NOS2 and nitrated amyloid-β (Fig. 4g, j) should contribute to the protection of synaptic plasticity, memory and behaviour.

These data are consistent with the hypothesis that amyloid-β-induced activation of the NLRP3 inflammasome enhances Alzheimer's disease progression by mediating a harmful chronic inflammatory tissue response. Inflammatory mediators that result from NLRP3 inflammasome activation are probably involved in mediating synaptic dysfunction, cognitive impairment and the restriction of beneficial microglial clearance functions. This key role of the NLRP3 inflammasome in amyloid-β-mediated inflammatory responses suggests that a therapeutic that blocks the activity of the NLRP3 inflammasome, or inflammasome-derived cytokines, might effectively interfere with the progression of Alzheimer's disease.

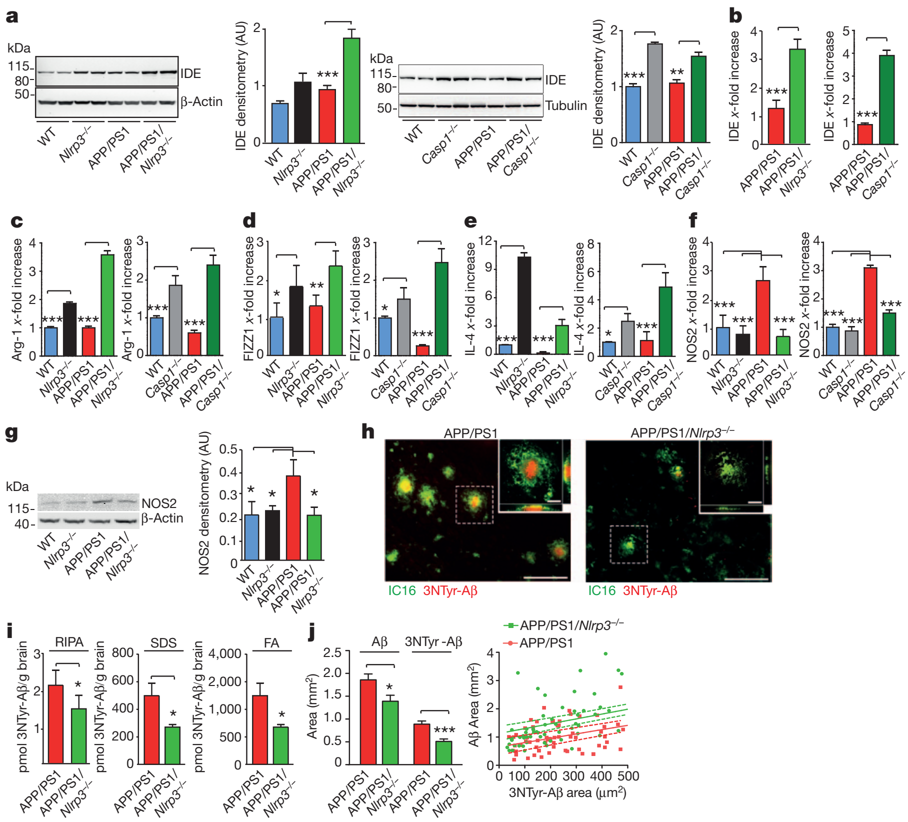


Figure 4 | NLRP3 gene deficiency conveys a M2 microglial phenotype, decreases NOS2 expression and strongly reduces 3NTyr-amyloid-β formation. **a**, Western blot detection of IDE in cerebral lysates of mice at 16 months of age. Quantification by densitometry is to the right of each western blot ($n = 5$, mean ± s.e.m., ANOVA, Tukey's test, $^{**}P < 0.01$, $^{***}P < 0.01$). **b**, Corresponding analysis of IDE gene transcription in caspase-1-deficient mice ($n = 5$ per group, mean ± s.e.m., Student's t-test, $^{***}P < 0.001$). **c**, Transcription of arginase-1 (Arg-1), (**d**) found in inflammatory zone-1 (FIZZ1), (**e**) IL-4 and (**f**) NOS2 at 16 months of age ($n = 5$, mean ± s.e.m., ANOVA, Tukey's *post hoc* test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.01$). **g**, Western blot detection and quantification of NOS2 in cerebral lysates from 16-month-old mice ($n = 5$, mean ± s.e.m., ANOVA, Tukey's *post hoc* test, $^{*}P < 0.05$). **h**, Representative brain sections were analysed by immunohistochemistry for nitrated amyloid-β (3NTyr-Aβ). **i**, ELISA detection of 3-NTyr-Aβ in RIPA, SDS and formic-acid extracts showed a robust reduction of 3NTyr-Aβ in APP/PS1/*Nlrp3*$^{-/-}$ mice at 16 months ($n = 4$–5, Student's t-test, $^{*}P < 0.05$). **j**, Left: cortical sections from 16-month-old mice were probed for 3NTyr-Aβ and amyloid-β using monoclonal antibody IC16. NLRP3 gene deficiency reduced both IC16-positive amyloid-β and 3NTyr-Aβ plaque size (85 plaques were assessed from each group of four mice, mean ± s.e.m., Student's t-test, $^{*}P < 0.05$, $^{***}P < 0.001$). Right: scatter plot of all plaques analysed by linear regression ($n = 4$ mice, 85 plaques per group; lines: linear regression analysis; dashed lines: 95% confidence intervals, $R^2 = 0.4920$ for APP/PS1 and $R^2 = 0.3884$ APP/PS1/*Nlrp3*$^{-/-}$ mice).

METHODS SUMMARY

Caspase-1 activation of human and mouse brain tissue was analysed by western blot of cleaved caspase-1. IL-1β was quantified by ELISA. Microglial ASC speck formation was detected by immunohistochemistry. All mice were on C57/Bl6 background, including WT, *Nlrp3*$^{-/-}$ (ref. 27), APP/PS1 (ref. 5), APP/PS1/*Nlrp3*$^{-/-}$, *Casp1*$^{-/-}$ (ref. 28) and APP/PS1/*Casp1*$^{-/-}$, and were analysed for cognitive function using the Morris water-maze test, the object recognition test and open field behavioural testing. Synaptic plasticity was determined by measuring LTP in acutely isolated hippocampal slices. Spine density was assessed by analysing mid-apical dendritic sections of pyramidal CA1 neurons. Cerebral amyloid-β load was determined by thioflavin S immunohistochemistry of serial sections. Sequential extraction of homogenized brains by radio-immunoprecipitation assay, sodium dodecyl sulphate buffer and formic acid was used to determine amounts of amyloid-β. Amyloid-β nitration was determined by ELISA and immunohistochemistry using specific antibodies against 3NTyr[10]-amyloid-β[25]. Western blot detection was used to analyse the protein amounts of APP, carboxyl-terminal fragments, amyloid-β, BACE1, IDE and NOS2. Inflammasome activation was confirmed by detection of ASC speck formation in microglia isolated from adult mouse. Microglial amyloid-β phagocytosis was determined after peripheral injection of methoxy-XO4, isolation of microglia and subsequent FACS analysis. Confirmatory immunocytochemistry was performed using antibody IC16 and the lysosomal marker LAMP2. Plaque morphology and microglial amyloid-β uptake were analysed by coimmunostaining with Iba-1, methoxy-XO4 and IC16. mRNA amounts of IDE, NEP, M1 and M2 markers were determined either from sorted microglia or from brain tissue by quantitative PCR.

Full Methods and any associated references are available in the online version of the paper.

Received 26 February; accepted 30 October 2012.
Published online 19 December 2012; corrected online 30 January 2013 (see full-text HTML version for details).

1. Prinz, M., Priller, J., Sisodia, S. S. & Ransohoff, R. M. Heterogeneity of CNS myeloid cells and their roles in neurodegeneration. *Nature Neurosci.* **14,** 1227–1235 (2011).
2. Lucin, K. M. & Wyss-Coray, T. Immune activation in brain aging and neurodegeneration: too much or too little? *Neuron* **64,** 110–122 (2009).
3. Halle, A. *et al.* The NALP3 inflammasome is involved in the innate immune response to amyloid-β. *Nature Immunol.* **9,** 857–865 (2008).
4. Martinon, F., Mayor, A. & Tschopp, J. The inflammasomes: guardians of the body. *Annu. Rev. Immunol.* **27,** 229–265 (2009).
5. Jankowsky, J. L. *et al.* Co-expression of multiple transgenes in mouse CNS: a comparison of strategies. *Biomol. Eng.* **17,** 157–165 (2001).
6. Bliss, T. V. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361,** 31–39 (1993).
7. Ho, V. M., Lee, J.-A. & Martin, K. C. The cell biology of synaptic plasticity. *Science* **334,** 623–628 (2011).
8. Walker, J. M. *et al.* Spatial learning and memory impairment and increased locomotion in a transgenic amyloid precursor protein mouse model of Alzheimer's disease. *Behav. Brain Res.* **222,** 169–175 (2011).
9. Heneka, M. T. & O'Banion, M. K. Inflammatory processes in Alzheimer's disease. *J. Neuroimmunol.* **184,** 69–91 (2007).
10. Lee, C. Y. D. & Landreth, G. E. The role of microglia in amyloid clearance from the AD brain. *J. Neural Transm.* **117,** 949–960 (2010).
11. Nalbantoglu, J. *et al.* Impaired learning and LTP in mice expressing the carboxy terminus of the Alzheimer amyloid precursor protein. *Nature* **387,** 500–505 (1997).
12. Chapman, P. F. *et al.* Impaired synaptic plasticity and learning in aged amyloid precursor protein transgenic mice. *Nature Neurosci.* **2,** 271–276 (1999).
13. Murray, C. A. & Lynch, M. A. Evidence that increased hippocampal expression of the cytokine interleukin-1β is a common trigger for age- and stress-induced impairments in long-term potentiation. *J. Neurosci.* **18,** 2974–2981 (1998).
14. El Khoury, J. *et al.* Ccr2 deficiency impairs microglial accumulation and accelerates progression of Alzheimer-like disease. *Nature Med.* **13,** 432–438 (2007).
15. Bamberger, M. E., Harris, M. E., McDonald, D. R., Husemann, J. & Landreth, G. E. A cell surface receptor complex for fibrillar β-amyloid mediates microglial activation. *J. Neurosci.* **23,** 2665–2674 (2003).
16. Hickman, S. E., Allison, E. K. & Khoury, J. E. Microglial dysfunction and defective β-amyloid clearance pathways in aging Alzheimer's disease mice. *J. Neurosci.* **28,** 8354–8360 (2008).
17. Heneka, M. T. *et al.* Locus ceruleus controls Alzheimer's disease pathology by modulating microglial functions through norepinephrine. *Proc. Natl Acad. Sci. USA* **107,** 6058–6063 (2010).
18. Shaftel, S. S. *et al.* Sustained hippocampal IL-1β overexpression mediates chronic neuroinflammation and ameliorates Alzheimer plaque pathology. *J. Clin. Invest.* **117,** 1595–1604 (2007).
19. Shaftel, S. S. *et al.* Chronic interleukin-1β expression in mouse brain leads to leukocyte infiltration and neutrophil-independent blood brain barrier permeability without overt neurodegeneration. *J. Neurosci.* **27,** 9301–9309 (2007).
20. Mildner, A. *et al.* Distinct and non-redundant roles of microglia and myeloid subsets in mouse models of Alzheimer's disease. *J. Neurosci.* **31,** 11159–11171 (2011).
21. Malito, E., Hulse, R. E. & Tang, W.-J. Amyloid β-degrading cryptidases: insulin degrading enzyme, neprilysin, and presequence peptidase. *Cell. Mol. Life Sci.* **65,** 2574–2585 (2008).
22. Leissring, M. A. *et al.* Enhanced proteolysis of β-amyloid in APP transgenic mice prevents plaque formation, secondary pathology, and premature death. *Neuron* **40,** 1087–1093 (2003).
23. Mawuenyega, K. G. *et al.* Decreased clearance of CNS β-amyloid in Alzheimer's disease. *Science* **330,** 1774 (2010).
24. Raes, G. *et al.* FIZZ1 and Ym as tools to discriminate between differentially activated macrophages. *Dev. Immunol.* **9,** 151–159 (2002).
25. Kummer, M. P. *et al.* Nitration of tyrosine 10 critically enhances amyloid β aggregation and plaque formation. *Neuron* **71,** 833–844 (2011).
26. Wang, Q., Rowan, M. J. & Anwyl, R. β-Amyloid-mediated inhibition of NMDA receptor-dependent long-term potentiation induction involves activation of microglia and stimulation of inducible nitric oxide synthase and superoxide. *J. Neurosci.* **24,** 6049–6056 (2004).
27. Kanneganti, T.-D. *et al.* Bacterial RNA and small antiviral compounds activate caspase-1 through cryopyrin/Nalp3. *Nature* **440,** 233–236 (2006).
28. Li, P. *et al.* Mice deficient in IL-1 β-converting enzyme are defective in production of mature IL-1β and resistant to endotoxic shock. *Cell* **80,** 401–411 (1995).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by the Dana Foundation (E.L.), the National Institutes of Health (E.L., D.T.G.) and the Deutsche Forschungsgemeinschaft (E.L., M.T.H.). We thank G. Nuñez and V. M. Dixit for providing anti-caspase-1 Abs. We thank B. De Strooper and L. Serneels for the BACE1 knockout mice and discussion. We also thank H. Jacobsen for the BACE1 transgenic mice.

Author Contributions M.T.H, M.P.K, A.S., A.D., S.S., A.V.-S., A.G., D.A., A.R., T.T. and E.L. performed experiments and analysed data, E.G. provided human samples and analysed data, A.H. was involved in study design and analysed data, E.L., M.T.H., M.K. and D.T.G. designed the study and wrote the paper. All authors discussed results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.L. (eicke.latz@uni-bonn.de), M.T.H. (michael.Heneka@ukb.uni-bonn.de) or D.T.G. (douglas.golenbock@umassmed.edu).

METHODS

Animals and ages. APP/PS1 transgenic animals[5] were obtained from The Jackson Laboratory (number 005864) on the C57BL/6 background. NLRP3-deficient animals[27] (Millennium Pharmaceuticals) were backbred onto C57BL/6 mice genotype to more than 99% C57BL/6, which was confirmed by microsatellite analysis. Caspase-1-deficient mice were generated by BASF[28]; these mice were subsequently obtained by M. Starnbach, who provided the mice for this study after 10 generations of backbreeding onto the C57BL/6 background. All mice were housed under standard conditions at 22 °C and a 12 h light:dark cycle with free access to food and water. Animal care and handling was performed according to the Declaration of Helsinki and approved by the local ethical committees. The following animal groups were analysed: WT, *Nlrp3*$^{-/-}$, APP/PS1, APP/PS1/*Nlrp3*$^{-/-}$, *Casp1*$^{-/-}$, APP/PS1/*Casp1*$^{-/-}$.

Human tissue samples. Post-mortem brain materials from histologically confirmed cases of Alzheimer's disease and age-matched controls who had died from non-neurological disease were from the Neurological Tissue Bank of the Biobank from the Hospital Clinic-Institut d'Investigacions Biomèdiques August Pi i Sunyer. Samples from patients with mild cognitive impairment and early-onset Alzheimer's disease were obtained from the Banner Health collection (http://www.bannerhealth.com). Ages and post-mortem sampling times were similar between controls and cases of mild cognitive impairment, early-onset Alzheimer's disease and Alzheimer's disease. Post-mortem times across all cases varied from 1.5 to 5 h. Patients were 75 ± 10 years old.

Behavioural phenotyping. For the Morris water-maze test, spatial memory testing was conducted in a pool consisting of a circular tank (Ø1 m) filled with opacified water at 24 °C. The water basin was dimly lit (20–30 lx) and surrounded by a white curtain. The maze was virtually divided into four quadrants, with one containing a hidden platform (15 × 15 cm), present 1.5 cm below the water surface. Mice were trained to find the platform, orienting by three extra maze cues placed asymmetrically as spatial references. They were placed into the water in a quasi-random fashion to prevent strategy learning. Mice were allowed to search for the platform for 40 s; if the mice did not reach the platform in the allotted time, they were placed onto it manually. Mice were allowed to stay on the platform for 15 s before the initiation of the next trial. After completion of four trials, mice were dried and placed back into their home cages. Mice trained four trials per day for eight consecutive days. For spatial probe trials, which were conducted 24 h after the last training session (day 9), the platform was removed and mice were allowed to swim for 30 s. The drop position was at the border between the third and fourth quadrant, with the mouse facing the wall at start. Data are given as the percentage of time spent in quadrant Q1, representing the quadrant where the platform had been located, and compared with the averaged time the animals spent in the remaining quadrants. In the afternoon of the same day, a visual cued testing was performed with the platform being flagged and new positions for the start and goal during each trial. All mouse movements were recorded by a computerized tracking system that calculated distances moved and latencies required for reaching the platform (Noldus, Ethovision 3.1). For open field exploration, mice were placed in the centre of the dimly lit (20–30 lx) chamber of the open field arena. Animal movements were tracked by an automatic monitoring system (Noldus Ethovision 3.1) for 5 min. The area was virtually divided into a centre (square with 40 cm edge lengths), a corridor (7.5 cm along the walls) and four corner squares (10 cm edge lengths), which partly overlapped with the corridor area. The time spent in each area, horizontal and vertical activity, frequency of urination and defaecation were monitored. The experiment was repeated on three consecutive days. The novel object recognition test was performed according to a previously established protocol with minor changes[29]. Briefly, the test procedure consisted of three sessions: habituation, training and retention. Each mouse was individually habituated to the open field arena. Habituation was allowed for 10 min. One day later, during the training session, two identical objects (object A) were placed into the two opposing corners of the centre area 30 cm apart from each other, and mice were allowed to explore the area and the objects for 10 min. The total time spent exploring both identical objects was recorded to examine place or object preference. Exactly 60 min or 1 day later, during the retention sessions, mice were placed back into the same arena in which one familiar (object A) and one novel object (object B) replacing the second object A were placed. Mice were then allowed to explore freely for 5 min and the time spent exploring each object was recorded. Exploration of the object was considered when the head of the animal was at least facing the object from a minimum distance of 1–2 cm or closer, but recording was cut as soon as mice turned their heads away from the previously investigated object. Time spent exploring the objects during trials was determined and is shown as discrimination ratio (novel object interaction/total object interaction). The arena and all objects were thoroughly cleaned with 70% ethanol solution after each trial.

Electrophysiology. For slice preparation, acute hippocampal transversal slices were prepared from 7- to 9-month-old WT, *Nlrp3*$^{-/-}$, APP/PS1 or combined APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice according to standard procedures. In brief, mice were deeply anaesthetized and the brain was quickly transferred into ice-cold carbogenated (95% O_2, 5% CO_2) artificial cerebrospinal fluid which contained 125 mM NaCl, 2 mM KCl, 1.25 mM NaH_2PO_4, 2 mM $MgCl_2$, 26 mM $NaHCO_3$, 2 mM $CaCl_2$ and 25 mM glucose. Hippocampi were dissected and cut into 400 μm transverse slices with a vibratome (Leica, VT1200S). Slices were maintained in carbogenated artificial cerebrospinal fluid at room temperature for at least 1.5 h before recording. Recordings were performed in a submerged recording chamber at 32 °C. For electrophysiology, after placing the slices in the submerged recording chamber, field excitatory postsynaptic potentials were recorded in stratum radiatum of CA1 region with a borosilicate glass micropipette (resistance 3–15 MΩ) filled with 3 M NaCl at a depth of 120–200 μm. Monopolar tungsten electrodes were used for stimulating the Schaffer collaterals at a frequency of 0.1 Hz. Stimulation was set to elicit a field excitatory postsynaptic potential with a slope of approximately 40% of maximum for LTP recordings and approximately 60% for long-term depression recordings. After 20 min baseline stimulation, LTP was induced by applying TBS. One burst consisted of four pulses at 100 Hz, repeated 10 times in a 200 ms interval. Three such bursts were used to induce LTP at 0.1 Hz. Basic synaptic transmission and presynaptic properties were analysed by input–output measurements and paired-pulse facilitation. The input–output measurements were performed by application of a defined value (paired-pulse facilitation) of current (25–250 μA in steps of 25 μA) and by adjusting the stimulus intensity to a certain current eliciting a fibre volley of desired voltage. Paired-pulse facilitation was measured by applying a pair of two stimuli in different inter-stimulus-intervals ranging from 10, 20, 40, 80 to 160 ms. Data were collected, stored and analysed with LABVIEW software (National Instruments). The initial slope of field excitatory postsynaptic potentials elicited by stimulation of the Schaffer collaterals was measured over time, normalized to baseline, which was the mean response of the 20 min before TBS application and plotted as average ± s.e.m. Parameters leading to an exclusion of single experiments were (1) an unstable baseline (variability more than ± 10%) or (2) a large population spike after TBS application producing an artefactually large LTP. Paired-pulse facilitation data were analysed by calculation of the ratio of the slope of the second field excitatory postsynaptic potential divided by the slope of the first one. All data were recorded and analysed in a blind fashion.

Tissue preparation. After completion of the behavioural testing, mice were deeply anaesthetized and transcardially perfused with 15 ml phosphate-buffered saline (PBS). The brains were removed from the skull. One hemisphere was frozen immediately for biochemical analysis and the other was either fixed in 4% paraformaldehyde or frozen over a mixture of dry ice and isopentane.

Brain protein extraction. Snap-frozen brain hemispheres were extracted as previously described[25]. Briefly, hemispheres were homogenized in PBS, 1 mM EDTA, 1 mM EGTA, 3 μl ml^{-1} protease inhibitor mix (Sigma). Homogenates were extracted in RIPA buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% NP40, 0.5% NaDOC, 0.1% SDS), centrifuged at 100,000*g* for 30 min and the pellet containing insoluble amyloid-β was solubilized in 2% SDS, 25 mM Tris-HCl, pH 7.5. In addition, the SDS-insoluble pellet was extracted with 70% formic acid in water. Formic acid was removed using a speed vac (Eppendorf) and the resulting pellet was solubilized in 200 mM Tris-HCl, pH 7.5.

Immunohistochemistry. Free-floating 40 μm serial sections were cut on a vibratome (Leica). Sections were stored in 0.1% NaN3, PBS. For immunohistochemistry, sections were treated with 50% methanol for 15 min. Then, sections were washed three times for 5 min in PBS and blocked in 3% BSA, 0.1% Triton X-100, PBS (blocking buffer) for 30 min followed by overnight incubation with the primary antibody in blocking buffer. Specificity controls were performed by staining with the secondary reagent and omission of the primary antibodies. Sections were washed three times in 0.1% Triton X-100, PBS and incubated with Alexa-488- or Alexa-594-conjugated secondary antibodies (1:500, Invitrogen) for 90 min, washed three times with 0.1% Triton X-100, PBS for 5 min. Sections were mounted using Immomount (Thermo). The following primary antibodies were used with respective concentrations: rabbit polyclonal anti-GFAP (1:800, Dako), rat monoclonal anti-mouse CD11b (MCA711, 1:400, Serotec), rabbit polyclonal anti-Iba1 (1:200, Wako), anti-ASC (1:200; AL177, AdipoGen), anti-NLRP3 (1:200; Cryo-2, AdipoGen) and anti-Fizz-1 (1:400 MAB1523, R&D Systems). Fluorescence microscopy was done on an Olympus BX61 equipped with a spinning disk unit or on an A1-MP laser scanning microscope (Nikon). Images were processed in CellˆP 3.5 (Olympus) or in NIS-elements 4 (Nikon). Alternatively, cryosections (20 μm) were fixed in 4% paraformaldehyde and immunostained using antibody IC16 (ref. 30) against human amyloid-β1-15 (1:400) or a specific antiserum against 3NTyr[10]-Aβ[25] following the above-described protocol.

Plaque histology. For thioflavin S staining, vibratome sections were rinsed in water, incubated in 0.01% thioflavin S in 50% ethanol and differentiated in 50% ethanol. Sections were analysed using a BX61 microscope equipped with a spinning-disk unit to achieve confocality (Olympus) or an A1-MP laser scanning microscope (Nikon). Image stacks were deconvoluted using CellˆP (Olympus). Quantitative

assessment of plaque areas was done using the MBF-ImageJ 1.43m software bundle (National Institutes of Health). In brief, total plaque number and amyloid-β area fraction were calculated using the software ImageJ 1.43m with plugins from the WCIF ImageJ collection. In particular, images were normalized and an automatic thresholding on the basis of the entropy of the histogram ('MaxEntropy') was used to identify the plaques. Pictures were converted to a binary and the 'fill holes' and 'watershed' algorithm were applied. Finally, plaque number, plaque area and average amyloid-β plaque size were calculated using the 'analyze particles' plugin of ImageJ. The amyloid-β area fraction was determined by dividing total plaque area by the area of the microscopic field. For staining plaques with methoxy-X04, sections were washed with PBS, incubated with 10 μM methoxy-XO4 in 50% DMSO/50% NaCl (0.9%), pH 12 for 10 min and washed twice with PBS before continuing with immunohistochemistry.

Protein blotting. Samples were separated by 4–12% NuPAGE (Invitrogen) using MES or MOPS buffer and transferred to nitrocellulose membranes. For caspase-1 blots, positive and negative controls were generated by precipitating supernatants from WT immortalized murine macrophages. For the negative control, cells were treated with 200 ng ml^{-1} lipopolysaccharide for 4 h. For the positive control, cells were treated with 200 ng ml^{-1} lipopolysaccharide for 3 h, followed by 10 μM nigericin for 1 h. APP and amyloid-β were detected using antibody 6E10 (Covance) and the C-terminal APP antibody 140 (CT15, gift from J. Walter). IDE was blotted using antibody PC730 (Calbiochem), caspase-1 using antibodies Casp1 clone 4B4.2.1 (gift from Genentech) and a caspase-1 antibody raised in rabbit (gift from G. Nuñez), neprilysin using antibody 56C6 (Santa Cruz), tubulin using antibody E7 (Developmental Studies Hybridoma Bank), BACE1 with antibody 2253 (ProSci), NOS2 using antibody 160862 (Cayman Chemicals), and β-actin using A2228 (Sigma) and 926-42212 (LI-COR Biosciences). Immunoreactivity was detected by enhanced chemiluminescence reaction (Millipore) or near-infrared detection (Odyssey, LI-COR). Chemoluminescence intensities were analysed using Chemidoc XRS documentation system (Biorad).

ELISA quantification of cerebral amyloid-β concentrations. Quantitative determination of amyloid-β was performed using an electrochemoluminescence ELISA for amyloid-β_{1-38}, amyloid-β_{1-40} and amyloid-β_{1-42} (Meso Scale Discovery). Signals were measured on a SECTOR Imager 2400 reader (Meso Scale Discovery). For ELISA determination of 3NTyr^{10}-Aβ, Mesocale L15XA 96-well plates were coated with 2 μg ml^{-1} of the monoclonal 3NTyr10-Aβ antibody 4A5E8 (own production) in PBS overnight at 4 °C. Plates were blocked with 5% blocker A (Meso Scale), 0.1% mouse gamma globulin (Rockland). SDS and formic-acid fractions from mouse brain were diluted in 1% blocker A, 0.1% mouse gamma globulin 1:25 and 1:100, respectively. Thirty-microlitre samples were incubated for 4 h at room temperature, washed with Tris wash buffer (Meso Scale) and incubated with 0.25 μg ml^{-1} MSD-tagged antibody 4G8 (Meso Scale) diluted in 1% blocker A, 0.1% mouse gamma globulin for 1 h at room temperature. Wells were washed with Tris wash buffer, and 150 μl of 2× read buffer (Meso Scale) was added.

ELISA quantification of cerebral IL-1β concentrations. Quantitative determination of IL-1β was performed using the MLB00C ELISA for the determination of murine IL-1β according to the protocol of the supplier (R&D Systems).

Quantitative PCR. RNA was extracted from brain tissues using the RNAeasy Micro Kit (Qiagen). Total RNA was quantified spectrophotometrically and reverse transcribed into complementary DNA using the RevertAid First Strand cDNA Synthesis kit (Fermentas) according to the manufacturer's instructions. Real-time quantitative PCR was performed using the StepOnePlus Real-Time PCR System (Applied Biosystems). The TaqMan gene expression assay and TaqMan universal PCR master mix (Applied Biosystems) was used for PCR amplification and real-time detection of PCR products. PCRs were done in 20 μl with 1 μl of the reverse transcribed product corresponding to 40 ng of total RNA, 1 μl of the gene expression assay mix and 10 μl of the master mix with the following temperature profile: 95 °C for 10 min and 45 cycles of 95 °C for 15 s and 60 °C for 1 min. mRNA expression values were normalized to the level of GAPDH expression. The following probes from LifeTechnologies were used: GAPDH (Mm99999915_g1), FIZZ-1 (Mm00445109_m1), NOS2 (Mm00440485_m1), Arg-1 (Mm00475988_m1), IL-4 (Mm00445259_m1), IDE (Mm00473077_m1); for BACE1 detection the following set of primers was used forward: ACAACCTGAGGGGAAAGTCC, reverse: TAC TACTGCCCGTGTCCACC. Analysis of the expression of the genes was performed using StepOne 2.2 software provided by Applied Biosystems.

Determination of amyloid-β-containing plaque-associated microglia. For the determination of amyloid-β-containing plaque-associated microglia, a double immunofluorescence staining for CD11b and amyloid-β used the antibodies described above. Fields of plaques were randomly selected in the cortex. Images were made in Cell-P with automatic illumination. The area of CD11b overlaying plaques was determined with the co-localization finder plugin in Image J 1.43m and corrected for total plaque area determined with the subroutine particle analysis after background subtraction equal for all images and binarization. Only plaques with a diameter smaller than 30 μm were included in the analysis. Per animal, a coverage of 50–200 plaques by microglia was determined. Animal number per group was five.

Assessment of microglial functions *in vivo*. For the *in vivo* amyloid-β phagocytosis assay, mice were intraperitonealy injected 3 h before being killed with 10 mg kg^{-1} methoxy-XO4 (provided by A. Verbruggen) in 50% DMSO/50% NaCl (0.9%), pH 12. Mice were perfused with ice-cold PBS and the brains were removed, chopped into pieces and incubated in HBSS, 10% FCS containing 0.144 mg ml^{-1} collagenase type IV for 1 h at 37 °C. Homogenization was achieved by pipetting gently up and down using a 19-gauge needle. The homogenate was filtered through a cell strainer (70 μm) and centrifuged at 155*g* and 4 °C for 10 min without a brake. The pellet was re-suspended in 9 ml 70% Percoll in PBS and underlayered with ice-cold 10 ml 37% Percoll in PBS and overlayered with 6 ml ice-cold PBS. The gradient was centrifuged at 800*g* and 4 °C for 25 min without a brake. Microglial cells were recovered from the 37/70% Percoll interphase, diluted with 3 vol. PBS and centrifuged at 880*g* and 4 °C for 25 min (Beckman Allegra) without a brake. The pellet containing the microglial cells was re-suspended in 200 μl PBS. For flow cytometry analysis, 50 μl of cells were diluted with 0.5 ml HBSS and centrifuged at 250*g* for 5 min at 4 °C. Binding of antibodies to Fc-receptors was prevented by adding 1 μg Fc-block and incubating for 10 min on ice. Cells were taken up in 50 μl of primary antibody mix (CD11b-APC (1:100, BioLegend, number 101212), CD45-FITC (1:100, eBioscience, number 11-0451), CD36-PE (1:100, eBioscience, number 12-0361)) and incubated for 30 min on ice. Cells were centrifuged at 250*g* for 5 min at 4 °C and re-suspended in 200 μl HBSS. For control and compensation, corresponding isotype control antibodies were used. Cells were measured on a FACSCanto II (BD Bioscience). For analysis, the CD11b$^+$ CD45$^+$ population was gated. WT mice injected with methoxy-XO4 were used to determine the methoxy-X04 threshold for non-phagocytosing cells, and unstained WT cells were used to determine background fluorescence. For FACS, microglia were stained with CD11b alone and sorted using a FACSDiVa cell sorter (BD Bioscience). To determine total brain methoxy-X04 fluorescence, APP/PS1 and APP/PS1/*Nlrp3*$^{-/-}$ were injected with methoxy-X04 as described above. After 3 h, brains were homogenized in PBS with 1 mM 4-(2-aminoethyl) benzenesulphonyl fluoride hydrochloride and 50 μl of the homogenate was measured at 368 nm excitation and 450 nm emission in a black 96-well plate using an infinite 200-plate reader (Tecan).

Immunocytochemistry of sorted microglia. A subset of microglial cells that were isolated according to the procedure described above were used to verify the uptake of methoxy-04 labelled amyloid-β by immunocytochemistry. Therefore, cells were brought onto glass slides by cytospin and subsequently fixed with 4% paraformaldehyde. Intracellular amyloid-β was visualized by double immunostaining for IC16 (ref. 30) and either CD11b (MCA711; AbSerotec) to detect microglial boundaries or LAMP2 using antibody Abl-93 (Developmental Studies Hybridoma Bank) to determine the intracellular localization of methoxy-X04 and IC16 positive amyloid-β. Inflammasome activation was visualized using the same cells and staining for CD11b and anti-ASC (AL177; AdipoGen).

DiOlistics and morphological analysis. Hippocampal neurons from WT, APP/PS1, APP/PS1/*Nlrp3*$^{-/-}$ and APP/PS1/*Casp1*$^{-/-}$ mice were labelled using DiOlistic on acute slices. Briefly, the mice were anaesthetized and decapitated, and the brain was quickly transferred into ice-cold carbogenated (95% O_2, 5% CO_2) artificial cerebrospinal fluid. Hippocampi were dissected and cut into 400 μm transversal slices with a vibratome (VT 1000S, Leica). Vibratome slices were immediately fixed in 4% PFA overnight at 4 °C. Tungsten particles (50 mg; 1.7 μm in diameter; Bio-Rad) were spread on a glass slide, and 100 μl of dye solution prepared by dissolving 3 mg of lipophilic dye DiI (Invitrogen) in 100 μl of methylenechloride (Sigma-Aldrich). The dried dye-coated particles were removed from the glass slide, re-suspended in 3 ml of distilled water and sonicated. The dye solution was subsequently diluted 1:100. To improve the bead attachment, the tube walls were precoated with a solution of PVP (polyvinyl-pyrrolidone) (stock: 0.05 mg ml^{-1} in ethanol; Bio-Rad), and the bullets were stored at room temperature. Dye-coated particles were delivered to the acute slices using a hand-held gene gun (Bio-Rad, Helios Gene Gun System). A membrane filter (3 μm; Millipore) was inserted between the gene gun and the preparation to prevent clusters of large particles from landing on the tissue. After shooting, slices were kept in PBS for 3 days at room temperature to allow dye diffusion. The slices were postfixed with 4% PFA, washed and mounted using an anti-fading water-based mounting medium (Biomeda). The spine density of pyramidal cells was measured for mid-apical dendrites. The selected dendrite segments were imaged using a LSM510 Meta confocal microscope (Zeiss) using a ×40 water-immersion objective and a zoom 4, and were z-sectioned at 0.5 μm. The number of spines was normalized per micrometre of dendritic length. We analysed statistics using GraphPad Prism 5.03. All data shown are presented as mean and s.e.m. The data obtained were compared between two different experimental conditions using a two-tailed Student's *t*-test. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

29. Bevins, R. A. & Besheer, J. Object recognition in rats and mice: a one-trial non-matching-to-sample learning task to study 'recognition memory'. *Nature Protocols* **1,** 1306–1311 (2006).

30. Jäger, S. *et al.* alpha-secretase mediated conversion of the amyloid precursor protein derived membrane stub C99 to C83 limits Aβ generation. *J. Neurochem.* **111,** 1369–1382 (2009).

doi:10.1038/nature11745

Regulation of mTORC1 by the Rag GTPases is necessary for neonatal autophagy and survival

Alejo Efeyan[1,2,3,4,5], Roberto Zoncu[1,2,3,4,5], Steven Chang[1,2,3,4,5], Iwona Gumper[6], Harriet Snitkin[6], Rachel L. Wolfson[1,2,3,4,5], Oktay Kirak[1]†, David D. Sabatini[6] & David M. Sabatini[1,2,3,4,5]

The mechanistic target of rapamycin complex 1 (mTORC1) pathway regulates organismal growth in response to many environmental cues, including nutrients and growth factors[1]. Cell-based studies showed that mTORC1 senses amino acids through the RagA–D family of GTPases[2,3] (also known as RRAGA, B, C and D), but their importance in mammalian physiology is unknown. Here we generate knock-in mice that express a constitutively active form of RagA (RagA^{GTP}) from its endogenous promoter. *RagA*$^{GTP/GTP}$ mice develop normally, but fail to survive postnatal day 1. When delivered by Caesarean section, fasted *RagA*$^{GTP/GTP}$ neonates die almost twice as rapidly as wild-type littermates. Within an hour of birth, wild-type neonates strongly inhibit mTORC1, which coincides with profound hypoglycaemia and a decrease in plasma amino-acid concentrations. In contrast, mTORC1 inhibition does not occur in *RagA*$^{GTP/GTP}$ neonates, despite identical reductions in blood nutrient amounts. With prolonged fasting, wild-type neonates recover their plasma glucose concentrations, but *RagA*$^{GTP/GTP}$ mice remain hypoglycaemic until death, despite using glycogen at a faster rate. The glucose homeostasis defect correlates with the inability of fasted *RagA*$^{GTP/GTP}$ neonates to trigger autophagy and produce amino acids for *de novo* glucose production. Because profound hypoglycaemia does not inhibit mTORC1 in *RagA*$^{GTP/GTP}$ neonates, we considered the possibility that the Rag pathway signals glucose as well as amino-acid sufficiency to mTORC1. Indeed, mTORC1 is resistant to glucose deprivation in *RagA*$^{GTP/GTP}$ fibroblasts, and glucose, like amino acids, controls its recruitment to the lysosomal surface, the site of mTORC1 activation. Thus, the Rag GTPases signal glucose and amino-acid concentrations to mTORC1, and have an unexpectedly key role in neonates in autophagy induction and thus nutrient homeostasis and viability.

The mechanistic target of rapamycin (mTOR) is a serine–threonine kinase that as part of mTORC1 regulates anabolic and catabolic processes required for cell growth and proliferation[1]. mTORC1 integrates signals that reflect the nutritional status of an organism and senses growth factors and nutrients through distinct mechanisms. Growth factors regulate mTORC1 through the PI3K/Akt/TSC1–TSC2 axis, whereas amino acids act through the Rag family of GTPases[2,3]. When activated, these GTPases recruit mTORC1 to the lysosomal surface, an essential step in mTORC1 activation[3,4]. Amino-acid concentrations regulate nucleotide binding to the Rag GTPases in a Ragulator- and vacuolar-type H^{+}-ATPase-dependent manner[4,5]. In the absence of amino acids, RagA (or RagB, which acts in an identical manner) is loaded with GDP, but becomes bound to GTP when amino acids are plentiful.

To study the physiological importance of the amino-acid-dependent activation of mTORC1, we generated knock-in mice that expressed a constitutively active form of RagA. We chose to manipulate RagA because, although highly similar to RagB, RagA is much more abundant and widely expressed than RagB in mice (Supplementary Fig. 1a). By a single nucleotide substitution in the RagA coding sequence, we replaced glutamine in position 66 with leucine, generating a RagA mutant (RagA^{Q66L}) (Supplementary Fig. 1b) that was, regardless of amino-acid concentrations, constitutively active, mimicking a permanent GTP-bound state[3,6] (hereafter referred to as RagA^{GTP}). We obtained mouse embryo fibroblasts (MEFs) from embryonic day (E)13.5 embryos and evaluated mTORC1 signalling upon amino-acid or serum deprivation. In *RagA*$^{+/+}$ and *RagA*$^{GTP/+}$ cells, deprivation of either amino acids (Fig. 1a) or serum (Supplementary Fig. 1c) suppressed mTORC1 activity, as determined by phosphorylation state of the mTORC1 substrates S6K1 and 4E-BP1. In contrast, in *RagA*$^{GTP/GTP}$ cells, mTORC1 activity was completely resistant to amino-acid withdrawal (Fig. 1a). However, regulation of PI3K activity by serum was intact, as reflected by Akt phosphorylation (Supplementary Fig. 1c). Interestingly, RagA protein was reduced in *RagA*$^{GTP/GTP}$ cells, but this was not a consequence of lower *RagA*^{GTP} messenger (mRNA) expression (Fig. 1b), supporting the existence of a negative feedback triggered by RagA activity. Nevertheless, the cells show the expected amino-acid-independent activation of mTORC1.

Cells lacking the TSC1–TSC2 tumour suppressor complex also have deregulated mTORC1 activity, as such cells maintain mTORC1 signalling in the absence of growth factors[1]. Unlike TSC1- or TSC2-deficient MEFs[7,8], *RagA*$^{GTP/GTP}$ MEFs have normal proliferation rates without accelerated senescence (Supplementary Fig. 1d). Furthermore, unlike TSC1- or TSC2-deficient embryos, which die at E11.5–E13.5, *RagA*$^{GTP/GTP}$ embryos were indistinguishable from *RagA*$^{+/+}$ embryos (Supplementary Fig. 1e), and fetuses were obtained and genotyped at term with the expected Mendelian ratios from heterozygous crosses. Thus, unlike with growth factor sensing, the inability of mTORC1 to sense amino-acid deprivation does not compromise survival during embryonic development, with its steady placental supply of nutrients.

Although apparently not deleterious during *in utero* development, constitutive RagA activity greatly impairs early postnatal survival. Heterozygous *RagA*$^{GTP/+}$ mice did not have any obvious phenotypic alteration, in agreement with the normal signalling observed in *RagA*$^{GTP/+}$ MEFs. However, no *RagA*$^{GTP/GTP}$ mice were obtained at weaning, and were usually found dead within 1 day postpartum in breeding cages. Neonatal death can stem from a variety of defects, so we obtained full-term E19.5 mice by Caesarean section and monitored them outside the breeding cage. Despite having a mild decrease in weight, *RagA*$^{GTP/GTP}$ neonates were barely distinguishable from control littermates (Fig. 1c, d), and histological analyses showed no abnormalities (Supplementary Fig. 1f).

To understand how constitutive RagA activity affects the regulation of mTORC1 by fasting, we compared the phosphorylation of S6 and 4E-BP1, established markers of mTORC1 activity, in tissues obtained from neonates at birth or fasted for 1 or 10 h. Interestingly, just 1 h of fasting was sufficient to inhibit mTORC1 strongly in *RagA*$^{+/+}$ and

[1]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Broad Institute of Harvard and Massachusetts Institute of Technology, Seven Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]David H. Koch Institute for Integrative Cancer Research at Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [5]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Department of Cell Biology, New York University School of Medicine, New York, New York 10016-6497, USA. †Present address: The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

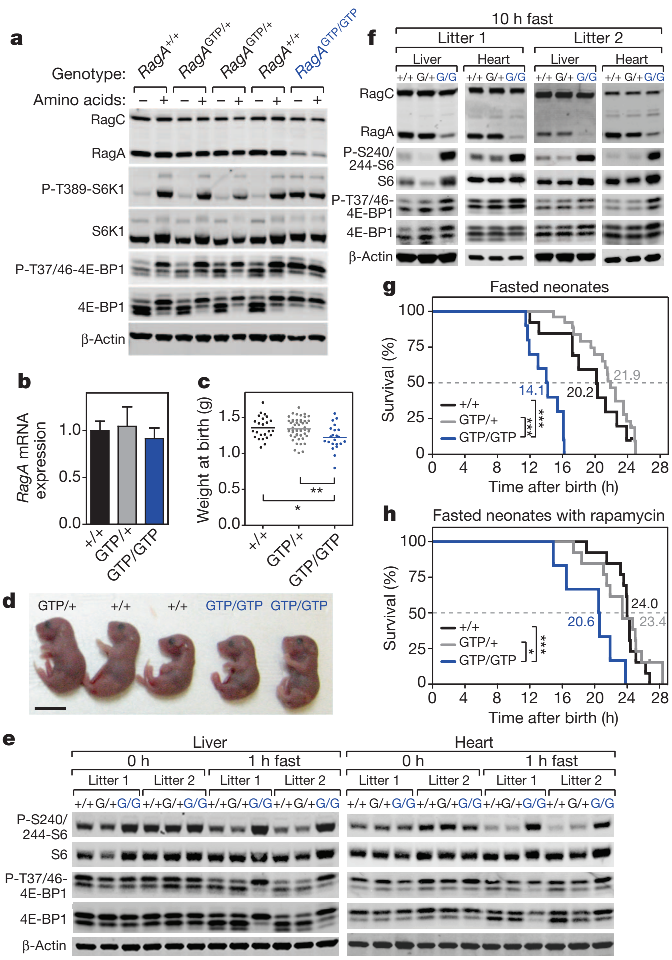


Figure 1 | Characterization of *RagA*$^{GTP/GTP}$ mice. a, MEFs of *RagA*$^{+/+}$, *RagA*$^{GTP/+}$ and *RagA*$^{GTP/GTP}$ genotypes were deprived of amino acids for 1 h and re-stimulated for 10 min. Whole-cell protein lysates were immunoblotted for indicated proteins. **b**, Total RNA was extracted from *RagA*$^{+/+}$ ($n = 3$), *RagA*$^{GTP/+}$ ($n = 3$) and *RagA*$^{GTP/GTP}$ ($n = 2$) MEFs and *RagA* mRNA expression determined by quantitative PCR (mean ± s.e.m.). **c**, Weights of *RagA*$^{+/+}$ ($n = 24$), *RagA*$^{GTP/+}$ ($n = 52$) and *RagA*$^{GTP/GTP}$ ($n = 22$) mice at birth (data are scatter dots, mean). **d**, Representative photographs of *RagA*$^{+/+}$, *RagA*$^{GTP/+}$ and *RagA*$^{GTP/GTP}$ neonates. Scale bar, 1 cm. **e**, Early suppression of mTORC1 signalling after birth was determined by immunoblotting of protein extracts from liver and heart of *RagA*$^{+/+}$ (+/+), *RagA*$^{GTP/+}$ (G/+) and *RagA*$^{GTP/GTP}$ (G/G) neonates immediately after Caesarean section (0 h) or after 1 h of fasting (1 h fast). **f**, Liver and heart extracts from *RagA*$^{+/+}$, *RagA*$^{GTP/+}$ and *RagA*$^{GTP/GTP}$ neonates fasted for 10 h were analysed by immunoblotting for the indicated proteins. **g**, Survival curve of fasted neonates. Neonates obtained by Caesarean section and resuscitated were fasted and their survival monitored (+/+: $n = 13$; G/+: $n = 26$; G/G: $n = 10$). **h**, Survival curve of fasted neonates treated with rapamycin. Neonates obtained by Caesarean section and resuscitated were fasted and their survival monitored (+/+: $n = 13$; G/+: $n = 13$; G/G: $n = 6$). Numbers indicate the median survival for each curve. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.005$.

RagA$^{GTP/+}$, but not *RagA*$^{GTP/GTP}$, neonates (Fig. 1e and Supplementary Fig. 1g); this difference persisted even after 10 h of fasting (Fig. 1f and Supplementary Fig. 1h). In contrast, Akt phosphorylation was modest at birth and decreased in mice of all genotypes (Supplementary Fig. 1g). As in MEFs, RagA protein was reduced in the tissues of *RagA*$^{GTP/GTP}$ mice, but this again was not due to reduced mRNA levels (Supplementary Fig. 1i). Collectively, these results indicate that constitutive RagA activity causes a profound defect in the response of mTORC1 to fasting.

To examine the consequences of this defect, we fasted neonates for longer periods, which showed that *RagA*$^{GTP/GTP}$ neonates have an accelerated time to death (approximately 14 h for *RagA*$^{GTP/GTP}$ compared with approximately 21 h in *RagA*$^{+/+}$ and *RagA*$^{GTP/+}$) (Fig. 1g). This was not the consequence of unappreciated developmental defects, as the treatment of pups at birth with the mTORC1 inhibitor rapamycin, which suppressed mTORC1 activity in all neonates (Supplementary Fig. 1j), significantly delayed the death of fasted *RagA*$^{GTP/GTP}$ neonates from approximately 14 h to 21 h; $P < 0.01$) (Fig. 1h). These data suggest that Rag-mediated regulation of mTORC1 is necessary for mice to adapt to and survive the starvation period that they endure between birth and feeding.

Consistent with this notion, analysis of blood glucose concentrations showed that fasted *RagA*$^{GTP/GTP}$ neonates suffer a profound defect in nutrient homeostasis. After 1 h of fasting, glycaemia dropped markedly in all neonates to below our 10 mg dl^{-1} limit of detection (Fig. 2a), but by 3–6 h the wild-type animals restored their blood glucose to near birth levels (approximately 40 mg dl^{-1}). In sharp contrast, in *RagA*$^{GTP/GTP}$ neonates, blood glucose concentrations never

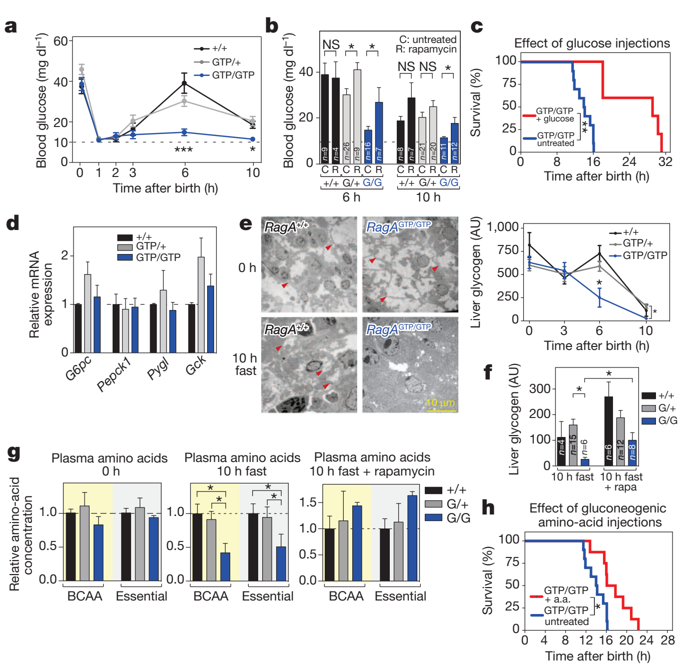


Figure 2 | Profound glucose homeostasis defect in *RagA*$^{GTP/GTP}$ mice. **a**, Glycaemia drop in *RagA*$^{+/+}$, *RagA*$^{GTP/+}$ and *RagA*$^{GTP/GTP}$ and recovery in fasted *RagA*$^{+/+}$ and *RagA*$^{GTP/+}$ but not in *RagA*$^{GTP/GTP}$ neonates (+/+: $n = 5$, 18, 4, 5, 9, 8; G/+: $n = 10$, 26, 10, 13, 26, 21; G/G: $n = 7$, 20, 9, 10, 16, 11, for 0, 1, 2, 3, 6 and 10 h, respectively; mean ± s.e.m.). **b**, Rapamycin significantly increases glycaemia in *RagA*$^{GTP/GTP}$ fasted for 6 and 10 h (mean ± s.e.m.). NS, not significant. **c**, Extension of survival by glucose injections in fasted *RagA*$^{GTP/GTP}$ neonates (untreated: $n = 10$; glucose: $n = 5$). **d**, Normal expression of genes involved in glucose metabolism in neonatal liver (+/+: $n = 2$, G/+: $n = 5$; G/G: $n = 4$; mean ± s.e.m.). **e**, Left: representative electron microscopy images showing abundant glycogen content in *RagA*$^{+/+}$ and *RagA*$^{GTP/GTP}$ hepatocytes before fasting (0 h, upper panels) and more pronounced glycogen depletion after 10 h of fasting (lower panels) in *RagA*$^{GTP/GTP}$ neonates. Right: quantification of hepatic glycogen content (+/+: $n = 5$, 3, 4, 4; G/+: $n = 11$, 7, 14, 15; G/G: $n = 6$, 4, 4, 6; for 0, 3, 6 and 10 h, respectively; mean ± s.e.m.; AU, arbitrary units). **f**, Partial rescue of hepatic glycogen content in *RagA*$^{GTP/GTP}$ neonates fasted for 10 h and treated with rapamycin (rapa) (mean ± s.e.m.). **g**, Quantification of neonatal plasma amounts of branched-chain (BCAA) and essential amino acids at birth (left), after 10 h fasting (middle) and after 10 h fasting with rapamycin treatment (right) (n for +/+, G/+ and G/G, respectively: $n = 4$, 5 and 4 for 0 h; $n = 4$, 4 and 3 for 10 h; $n = 2$, 5 and 3 for rapamycin; mean ± s.d.). Values are expressed relative to *RagA*$^{+/+}$ amounts at each time point. **h**, Extension of survival by injection of a combination of gluconeogenic amino acids (a.a.) in fasted *RagA*$^{GTP/GTP}$ neonates (untreated: $n = 10$; amino acids: $n = 8$). $^{*}P < 0.05$; $^{**}P < 0.01$; ns, $P > 0.05$.

recovered and remained at approximately 10 mg dl^{-1} or lower until death (Fig. 2a). Consistent with its rescue of the accelerated lethality of the *RagA*$^{GTP/GTP}$ neonates during fasting (Fig. 1h), rapamycin administration partly reversed their defect in blood glucose concentrations (Fig. 2b). Moreover, injections of glucose prolonged the lifespan of fasted *RagA*$^{GTP/GTP}$ mice (Fig. 2c), arguing that a lack of glucose has a causal role in their compromised survival.

Because the inability to generate glucose from glycogen can cause perinatal lethality[9], we initially proposed that the *RagA*$^{GTP/GTP}$ neonates had a glycogen metabolism defect. However, *RagA*$^{GTP/GTP}$ neonates did not have defects in the protein or mRNA levels of the key enzymes of glycogen metabolism (Fig. 2d and Supplementary Fig. 2a). Moreover, at birth *RagA*$^{GTP/GTP}$ neonates had normal amounts of hepatic glycogen, which, upon fasting, they consumed at a faster rate than *RagA*$^{+/+}$ and *RagA*$^{GTP/+}$ animals (Fig. 2e), suggesting not a defect in its breakdown but rather accelerated use secondary to hypoglycaemia. As with other characteristics of *RagA*$^{GTP/GTP}$ mice, rapamycin administration partly restored their hepatic glycogen (Fig. 2f).

We also considered defects in gluconeogenesis or the availability of gluconeogenic substrates as potential reasons for the inability of *RagA*$^{GTP/GTP}$ neonates to restore blood glucose concentrations upon fasting. Here too the *RagA*$^{GTP/GTP}$ neonates did not have aberrations in the expression levels of the relevant enzymes (Fig. 2d). However, after a fast for 10 h, *RagA*$^{GTP/GTP}$ neonates did have significantly lower levels of plasma amino acids compared with *RagA*$^{+/+}$ and *RagA*$^{GTP/+}$ littermates (Fig. 2g and Supplementary Fig. 2b). Because murine neonates are born without significant fat stores[10], lipid mobilization cannot serve as a substrate for *de novo* glucose production. Moreover, lactate, another substrate for gluconeogenesis, was not reduced in *RagA*$^{GTP/GTP}$ neonates (Supplementary Fig. 2c), arguing for a specific reduction in amino-acid substrates. As with glucose amounts and glycogen stores (Fig. 2b, f), rapamycin administration reversed the decrease in amino-acid concentrations in *RagA*$^{GTP/GTP}$ neonates (Fig. 2g). Furthermore, injection of a mix of gluconeogenic amino acids, which can contribute to gluconeogenesis but not protein synthesis, delayed the onset of death of *RagA*$^{GTP/GTP}$ neonates (Fig. 2h), and injection of just alanine to fasted neonates provoked a significant increase in glycaemia (Supplementary Fig. 2d). These data are consistent with the glucose homeostasis defect of the fasted *RagA*$^{GTP/GTP}$ neonates being a consequence of reduced circulating amino acids, which leads to lower *de novo* glucose production and plasma levels, and accelerated death.

Several properties of the *RagA*$^{GTP/GTP}$ mice are reminiscent of autophagy-deficient mice[11,12], including the reduction in plasma amino acids and lifespan upon fasting, as well as the slightly lower birth weight. Although mTORC1 negatively regulates autophagy[13], and amino-acid concentrations are regulators of autophagy in rats[14], many mTORC1-dependent and mTORC1-independent autophagy regulators exist[15]. Hence, we wondered if perturbing just one of the several inputs to mTORC1 could exert a dominant effect in the physiological regulation of autophagy.

Quantitative electron microscopy of livers from *RagA*$^{+/+}$ neonates fasted for 1 h showed abundant autophagosomes, characterized by double limiting membranes (Fig. 3a and Supplementary Fig. 3a). Autophagosomes rapidly mature into single-membrane autophagolysosomes, so these were also found in *RagA*$^{+/+}$ livers (Fig. 3a and Supplementary Fig. 3a), albeit the ratio of autophagosomes to autophagolysosomes was high. Both autophagic vacuoles were rarely observed in fasted *RagA*$^{GTP/GTP}$ littermates (Fig. 3a). Similar results were obtained when skeletal muscle was analysed (Fig. 3a).

Even after 10 h of fasting, the autophagy defect in the livers of *RagA*$^{GTP/GTP}$ neonates persisted, as detected by the reduced cleavage of LC3B and degradation of p62 (Fig. 3b), which was increased by administration of rapamycin (Supplementary Fig. 3b). Biochemical analyses for these markers in skeletal and cardiac muscles from *RagA*$^{GTP/GTP}$ neonates after 1 and 2 h of fasting were also consistent with impaired autophagy (Fig. 3c). Cells in culture mirrored the *in vivo* findings (Fig. 3d), and these results were confirmed by detection of LC3B localization using fluorescence microscopy in amino-acid-starved cells (Fig. 3e and Supplementary Fig. 3c). Consistently, phosphorylation of the autophagy activator ULK-1, a direct substrate of mTORC1 that was suppressed in *RagA*$^{+/+}$ MEFs upon amino-acid withdrawal, remained high in *RagA*$^{GTP/GTP}$ cells (Fig. 3d). In addition, we looked at the transcription factor TFEB, which upregulates genes involved in lysosomal biogenesis and autophagy, but is excluded from the nucleus when phosphorylated by mTORC1 (refs 16, 17). Upon amino-acid deprivation, TFEB localized to the nuclei of *RagA*$^{+/+}$ but not *RagA*$^{GTP/GTP}$ MEFs (Fig. 3f and Supplementary Fig. 3d). This result was mirrored by the decreased expression of TFEB transcriptional targets (Supplementary Fig. 3d).

Serum withdrawal, which inhibits mTORC1 in a Rag-independent fashion, suppressed mTORC1 activity and triggered autophagy in MEFs of all genotypes (Supplementary Fig. 3e), indicating that constitutive

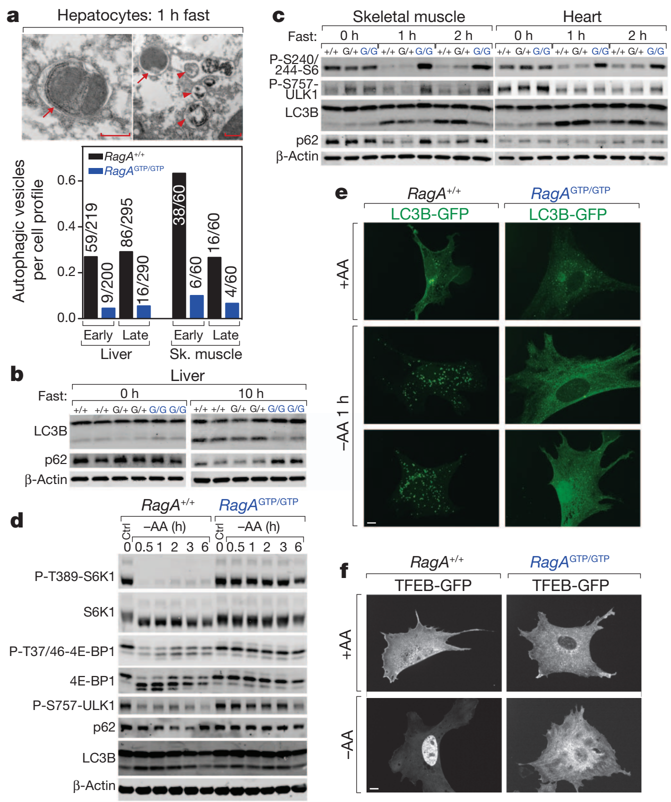


Figure 3 | Impaired autophagy in *RagA*$^{GTP/GTP}$ neonates. a, Top: representative micrographs of autophagosomes and autophagolysosomes in hepatocytes from *RagA*$^{+/+}$ neonates fasted for 1 h. Typical autophagosome with a double limiting membrane (arrows); autophagosome and several autolysosomes (arrowheads). Scale bars, 5 μm. Bottom: frequency of the two types of organelle (early: autophagosomes; late: autophagolysosomes) detected in cell profiles of hepatocytes and skeletal myocytes from *RagA*$^{+/+}$ and *RagA*$^{GTP/GTP}$ neonates. **b**, Protein extracts from livers of neonates at Caesarean section (0 h) and fasted for 10 h were immunoblotted for autophagy markers LC3B and p62. **c**, Protein extracts from skeletal muscle and heart from neonates at Caesarean section (0 h), fasted for 1 and 2 h, were immunoblotted for indicated proteins. **d**, Triggering of autophagy by amino-acid withdrawal in MEFs. MEFs were deprived of amino acids for the indicated time points, whole-cell protein extracts were obtained and mTORC1 activity and autophagic activity determined by immunoblotting. **e**, Recombinant LC3B–GFP was expressed in *RagA*$^{+/+}$ and *RagA*$^{GTP/GTP}$ MEFs and LC3B localization, in the presence and absence of amino acids, monitored by fluorescence microscopy. LC3B–GFP (green fluorescent protein) clustering, indicative of autophagy, was observed in amino acid-starved *RagA*$^{+/+}$ but not *RagA*$^{GTP/GTP}$ MEFs. Scale bar, 10 μm. **f**, Localization of recombinant TFEB–GFP was determined in MEFs as in **e**. Nuclear (active) TFEB was observed in *RagA*$^{+/+}$ MEFs upon amino-acid withdrawal.

RagA activity does not block autophagy induction by all signals. Thus, despite the multitude of pathways that regulate autophagy[15], Rag GTPase activity upstream of mTORC1 is a major regulator of autophagy *in vivo* during the perinatal period.

Maintenance of mTORC1 activity requires the simultaneous presence of growth factors, amino acids and glucose[1]. We found that after just 1 h of fasting, both plasma amino-acid and glucose concentrations were reduced in neonates of all genotypes (Fig. 2a and Supplementary Fig. 4a). The drop in nutrients correlated with a strong inhibition of mTORC1 activity in $RagA^{+/+}$ and $RagA^{GTP/+}$, but not $RagA^{GTP/GTP}$, neonates (Fig. 1e). Thus, despite a profound hypoglycaemic state, mTORC1 activity remained high in fasted $RagA^{GTP/GTP}$ neonates, a puzzling result given that the Rag GTPases are thought to have a specialized role in amino-acid sensing. These observations led us to consider that the Rag GTPases participate in the direct sensing of glucose concentrations, in addition to their established role in amino-acid sensing. A well-established link between low glucose (but not amino acids (Supplementary Fig. 4b)) and mTORC1 inhibition is the AMP-activated protein kinase (AMPK). However, in MEFs lacking AMPK-α1 and -α2 (AMPK-DKO), mTORC1 activity was still repressed upon glucose deprivation, albeit less prominently than in wild-type MEFs (Fig. 4a). This indicates that an AMPK-independent pathway of mTORC1 inhibition exists, as shown recently in the context of metformin treatment[18]. Compared with control cells, mTORC1 signalling was largely resistant to glucose deprivation in $RagA^{GTP/GTP}$ MEFs (Fig. 4b and Supplementary Fig. 4c, e) and HEK-293T cells expressing $RagB^{GTP}$ (Supplementary Fig. 4d, e). It is unlikely that glucose indirectly inhibits mTORC1 by preventing amino-acid transport, because amino-acid esters, which freely enter cells and substitute for native amino acids in mTORC1 activation[5], did not substitute for glucose (Supplementary Fig. 4f). Moreover, intracellular amino-acid concentrations were only marginally affected in cells deprived of glucose (Supplementary Fig. 4g). In addition, like AMPK-deficient cells[19,20], $RagA^{GTP/GTP}$ cells had enhanced sensitivity to long-term

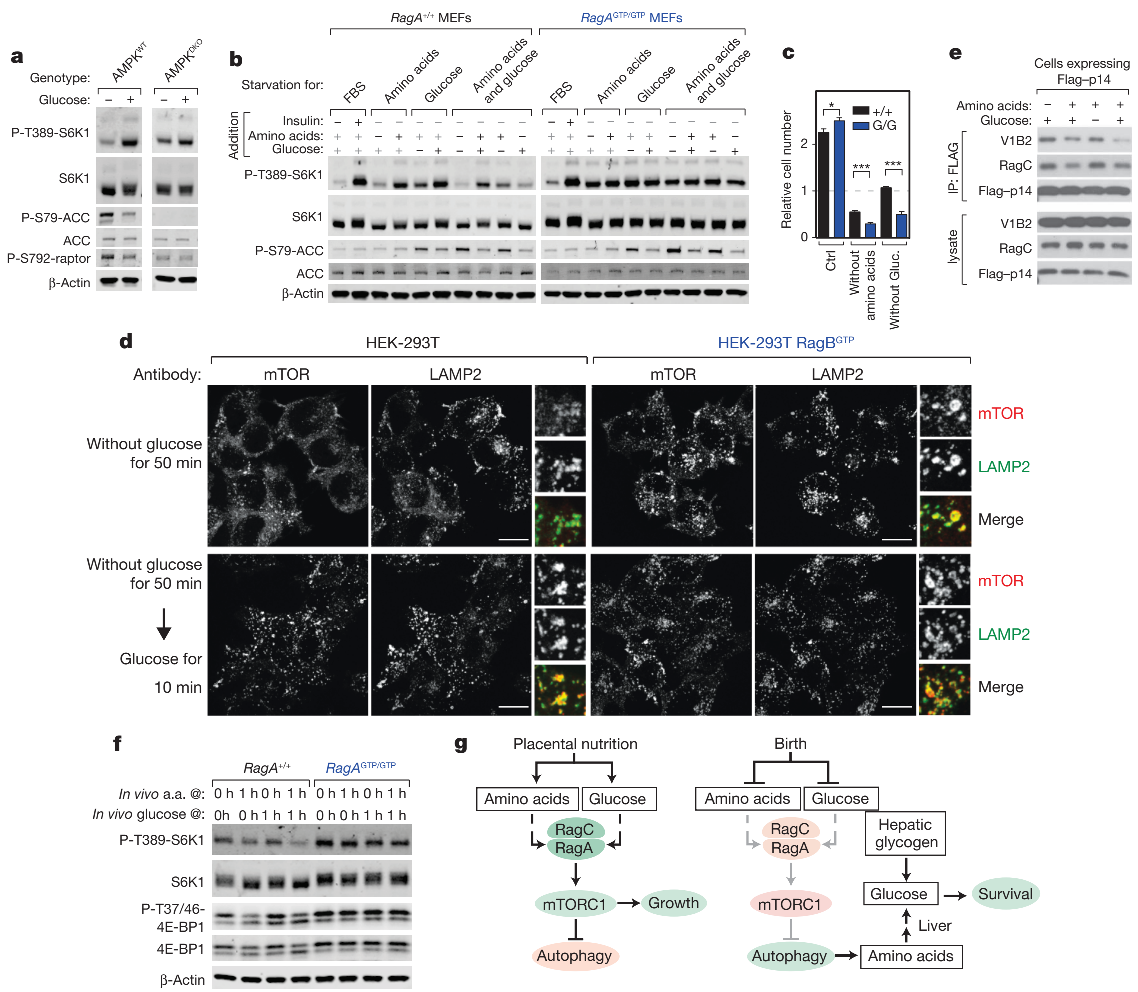


Figure 4 | The Rag GTPases mediate inhibition of mTORC1 by glucose deprivation. **a**, AMPKα1/α2 double knockout (DKO) and wild-type (WT) MEFs were deprived of glucose for 1 h and re-stimulated for 10 min. Whole-cell extracts were immunoblotted for the indicated proteins. **b**, Immortalized MEFs of the indicated genotypes were deprived of growth factors, glucose, amino acids or glucose and amino acids for 1 h and re-stimulated with glucose and/or amino acids for 10 min. Whole-cell lysates were immunoblotted for the indicated proteins. **c**, $RagA^{+/+}$ and $RagA^{GTP/GTP}$ immortalized MEFs were deprived of glucose or amino acids and surviving cells quantified in triplicate after 48 h. Cell number is indicated relative to cell number at the start of the treatment; mean ± s.d.; $***P < 0.005$. **d**, mTOR localization as detected by immunofluorescence. In HEK-293T cells, glucose deprivation causes mTOR to localize to diffuse small puncta throughout the cytoplasm. Re-addition of glucose leads to mTOR shuttling to the lysosomal surface, co-localizing with the lysosomal protein Lamp2. HEK-293T-$RagB^{GTP}$ cells show mTOR localized at the lysosomal surface, regardless of glucose concentrations. Scale bars, 10 μm. **e**, Glucose and amino acids affect the binding of the v-ATPase to the Ragulator complex. HEK-293T expressing Flag-p14 was deprived of glucose or amino acids for 90 min and re-stimulated for 20 min. Protein extracts and immunoprecipitates were immunoblotted for the indicated proteins. **f**, $RagA^{+/+}$ and $RagA^{GTP/GTP}$ primary MEFs were cultured for 1 h in media with the glucose and amino-acid concentrations measured in neonates at birth (0 h) or after fasting for 1 h (1 h) and whole-cell protein extracts were analysed by immunoblotting. **g**, Proposed model for constitutive RagA-induced neonatal lethality. Green and red boxes indicate active and inactive protein or process, respectively.

glucose-deprivation-induced death (Fig. 4c). Constitutive RagA activity does not block AMPK action as aminoimidazole carboxamide ribonucleotide (AICAR), an AMPK activator, inhibited mTORC1 in cells of all genotypes (Supplementary Fig. 4h). In addition, AMPK activity, as monitored by acetyl-CoA carboxylase phosphorylation, was induced to similar amounts in glucose-deprived *RagA*$^{+/+}$ and *RagA*$^{GTP/GTP}$ cells (Fig. 4b and Supplementary Fig. 4c), but absent in AMPK-null cells (Fig. 4a). Another cellular nutrient sensor is GCN2 (ref. 21), but although it was regulated by amino acids, it was not by glucose; also, loss of GCN2 did not affect the inhibition of mTORC1 caused by amino-acid or glucose starvation (Supplementary Fig. 4i).

Amino acids promote the Rag-dependent translocation of mTORC1 to the lysosomal surface, a necessary event for its activation[4]. Interestingly, glucose deprivation, like that of amino acids (Supplementary Fig. 4j), rendered mTORC1 diffusely localized in the cytoplasm of HEK-293T cells and, within minutes of glucose re-addition, mTORC1 re-clustered at lysosomes (Fig. 4d). However, in HEK-293T cells expressing RagB^{GTP} and in *RagA*$^{GTP/GTP}$ MEFs, mTORC1 localized at the lysosomal surface regardless of glucose concentrations (Fig. 4d and Supplementary Fig. 4k). The lysosomal v-ATPase, necessary for the Rag-dependent activation of mTORC1 by amino acids, engages in amino-acid-sensitive interactions with the Ragulator[5], and we found that glucose also regulates the binding of the v-ATPase to Ragulator (Fig. 4e), suggesting a shared regulatory mechanism. Finally, when amino-acid and glucose concentrations at birth and after 1 h neonatal fasting were reproduced in the *in vitro* culture medium, mTORC1 activity was suppressed in *RagA*$^{+/+}$ but not in *RagA*$^{GTP/GTP}$ cells placed under the 1 h fasting conditions (Fig. 4f). Hence, we propose that the Rag GTPases are a 'multi-input nutrient sensor', upon which amino acids and glucose converge, in a v-ATPase-dependent manner, upstream of mTORC1.

Altogether, our results support a chain of events that start with the interruption of maternal nutrient supply at birth, which inhibits mTORC1 presumably by converging negative inputs from profound hypoglycaemia and a drop in plasma amino acids, in a Rag-dependent fashion. During the period between birth and suckling, mTORC1 inhibition triggers autophagy, which generates the amino acids used to sustain plasma glucose concentrations through gluconeogenesis. Constitutive RagA activity prevents mTORC1 inhibition, leading to defective autophagy and, thus, insufficient amino-acid production. The lower levels of gluconeogenic amino acids reduce hepatic generation of glucose, which accelerated glycogen breakdown fails to compensate, ultimately leading to hypoglycaemia, energetic exhaustion and accelerated neonatal death (Fig. 4g). Thus, the Rag GTPases have a critical role in nutrient sensing by mTORC1 and in neonatal survival during fasting.

METHODS SUMMARY

All animal studies and procedures were approved by the Massachusetts Institute of Technology Institutional Animal Care and Use Committee. To target the *RagA* locus, we generated a construct consisting of a transcriptional STOP cassette containing the hygromycin resistance gene, flanked by *loxP* sites (*loxP-PGK-Hyg*-STOP-*loxP*)[22] and placed at 5′ of the *RagA* exon. A RagA activating mutation (Q66L) was generated by an A-to-T substitution in position +197 in the *RagA* exon by site-directed mutagenesis. Chimaeras were crossed to CMV-Cre transgenic mice to allow expression of the *RagA*^{Q66L} allele. Neonates were obtained by Caesarean section and placed in a humidified chamber at 30 °C and fasted. Subcutaneous injections of rapamycin were performed after Caesarean section, and those of glucose or a mix of gluconeogenic amino acids were performed after Caesarean section and at 3–6 h intervals. Neonatal plasma amino acids were quantified with an Acquity UPLC system (Waters), and hepatic glycogen content as described[23]. For statistical analyses, a log-rank Mantel–Cox method was used for Kaplan–Meier survival curves, and non-parametric *t*-tests and 2×2 χ^2 tests were performed as stated in the legends to the figures.

Full Methods and any associated references are available in the online version of the paper.

Received 15 February; accepted 5 November 2012.
Published online 23 December 2012.

1. Zoncu, R., Efeyan, A. & Sabatini, D. M. mTOR: from growth signal integration to cancer, diabetes and ageing. *Nature Rev. Mol. Cell Biol.* **12,** 21–35 (2010).
2. Kim, E., Goraksha-Hicks, P., Li, L., Neufeld, T. P. & Guan, K. L. Regulation of TORC1 by Rag GTPases in nutrient response. *Nature Cell Biol.* **10,** 935–945 (2008).
3. Sancak, Y. *et al.* The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320,** 1496–1501 (2008).
4. Sancak, Y. *et al.* Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids. *Cell* **141,** 290–303 (2010).
5. Zoncu, R. *et al.* mTORC1 senses lysosomal amino acids through an inside-out mechanism that requires the vacuolar H-ATPase. *Science* **334,** 678–683 (2011).
6. Hirose, E., Nakashima, N., Sekiguchi, T. & Nishimoto, T. RagA is a functional homologue of *S. cerevisiae* Gtr1p involved in the Ran/Gsp1-GTPase pathway. *J. Cell Sci.* **111,** 11–21 (1998).
7. Kwiatkowski, D. J. *et al.* A mouse model of TSC1 reveals sex-dependent lethality from liver hemangiomas, and up-regulation of p70S6 kinase activity in Tsc1 null cells. *Hum. Mol. Genet.* **11,** 525–534 (2002).
8. Zhang, H. *et al.* Loss of Tsc1/Tsc2 activates mTOR and disrupts PI3K-Akt signaling through downregulation of PDGFR. *J. Clin. Invest.* **112,** 1223–1233 (2003).
9. Scheuner, D. *et al.* Translational control is required for the unfolded protein response and in vivo glucose homeostasis. *Mol. Cell* **7,** 1165–1176 (2001).
10. Birsoy, K. *et al.* Analysis of gene networks in white adipose tissue development reveals a role for ETS2 in adipogenesis. *Development* **138,** 4709–4719 (2011).
11. Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432,** 1032–1036 (2004).
12. Komatsu, M. *et al.* Impairment of starvation-induced and constitutive autophagy in *Atg7*-deficient mice. *J. Cell Biol.* **169,** 425–434 (2005).
13. Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. Autophagy fights disease through cellular self-digestion. *Nature* **451,** 1069–1075 (2008).
14. Mortimore, G. E. & Schworer, C. M. Induction of autophagy by amino-acid deprivation in perfused rat liver. *Nature* **270,** 174–176 (1977).
15. Kroemer, G., Marino, G. & Levine, B. Autophagy and the integrated stress response. *Mol. Cell* **40,** 280–293 (2010).
16. Roczniak-Ferguson, A. *et al.* The transcription factor TFEB links mTORC1 signaling to transcriptional control of lysosome homeostasis. *Sci. Signal.* **5,** ra42 (2012).
17. Settembre, C. *et al.* A lysosome-to-nucleus signalling mechanism senses and regulates the lysosome via mTOR and TFEB. *EMBO J.* **31,** 1095–1108 (2012).
18. Kalender, A. *et al.* Metformin, independent of AMPK, inhibits mTORC1 in a rag GTPase-dependent manner. *Cell Metab.* **11,** 390–401 (2010).
19. Choo, A. Y. *et al.* Glucose addiction of TSC null cells is caused by failed mTORC1-dependent balancing of metabolic demand with supply. *Mol. Cell* **38,** 487–499 (2010).
20. Shaw, R. J. *et al.* The tumor suppressor LKB1 kinase directly activates AMP-activated kinase and regulates apoptosis in response to energy stress. *Proc. Natl Acad. Sci. USA* **101,** 3329–3335 (2004).
21. Proud, C. G. eIF2 and the control of cell physiology. *Semin. Cell Dev. Biol.* **16,** 3–12 (2005).
22. Guerra, C. *et al.* Tumor induction by an endogenous *K-ras* oncogene is highly dependent on cellular context. *Cancer Cell* **4,** 111–120 (2003).
23. Lo, S., Russell, J. C. & Taylor, A. W. Determination of glycogen in small tissue samples. *J. Appl. Physiol.* **28,** 234–236 (1970).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Sabatini laboratory for suggestions, and A. Hutchins for technical assistance. We thank R. Shaw for providing the AMPK-DKO MEFs, D. Ron for the GCN2-KO MEFs and M. Barbacid for the transcriptional STOP cassette. This work was supported by grants from the National Institutes of Health (R01 CA129105, R01 CA103866 and R37 AI047389) and awards from the American Federation for Aging, Starr Foundation, Koch Institute Frontier Research Program, and the Ellison Medical Foundation to D.M.S., fellowships from the Human Frontiers Science Program to A.E., and the Jane Coffin Childs Memorial Fund for Medical Research and the LAM Foundation to R.Z. D.M.S. in an investigator of Howard Hughes Medical Institute.

Author Contributions A.E. and D.M.S. conceived the project. A.E. designed and performed most experiments with input from D.M.S, and assistance from S.C., R.L.W. and O.K. R.Z. performed experiments and participated in discussion of the results. I.G., H.S. and D.D.S. performed electron microscopy experiments and interpretations. D.D.S. helped with discussion and interpretation of results. A.E. wrote and D.M.S. edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.M.S. (sabatini@wi.mit.edu).

METHODS

Generation of *RagA*^{GTP} mice. All animal work was performed in accordance with the Massachusetts Institute of Technology Committee on Animal Care. To target the *RagA* locus, we generated a construct consisting of a 4-kilobase 5′ homology arm upstream of the *RagA* gene, a transcriptional STOP cassette containing a the hygromycin resistance gene, flanked by *loxP* sites (*loxP-PGK-Hyg*-STOP-*loxP*)[22] placed at the 5′ end of the *RagA* exon, followed by a 3-kilobase 3′ homology arm downstream of *RagA* genomic sequence. For cloning purposes, a NotI restriction (GCGGCCGC) site was inserted into the 5′ homology arm, replacing the GGCGACGC sequence located 17 nucleotides upstream of the *RagA* ATG translation start codon. The *loxP-PGK-Hyg*-STOP-*loxP*, previously cloned in the pMeca plasmid, was excised with NotI and inserted in the *RagA* construct. An A-to-T substitution in position +197 in *RagA* exon was performed by site-directed mutagenesis. This mutation translates into a Q66L amino-acid substitution that renders a RagA protein that is constitutively active. See also Supplementary Fig. 1b. The construct was inserted into pPGKNeo.F2L2.DTA, linearized with XmaI and electroporated into male embryonic stem cells of mixed 129Sv/C57B6 background. Embryonic stem cell colonies were picked and identified by Southern blot and confirmed by PCR amplification of specific insertion products. Positive embryonic stem cell clones were then injected into blastocysts and transferred into pseudo-pregnant females to obtain chimaeric mice. Male chimaeras were crossed to CMV-Cre transgenic females of C57BL/6J background, resulting in excision of the transcriptional STOP cassette and allowing expression of the *RagA*^{Q66L} allele, then intercrossed.

Preparation of MEFs. MEFs from E13.5 embryos of *RagA*$^{+/+}$, *RagA*$^{GTP/+}$ and *RagA*$^{GTP/GTP}$ genotype were prepared by chemical digestion followed by mechanical disaggregation. For spontaneous immortalization, MEFs were re-plated every 3 days until senescent. Spontaneously proliferating cells eventually arose after a senescent phase. AMPKα1/α2 DKO immortalized MEFs and matched wild-type MEFs were provided by R. Shaw; D. Ron provided the GCN2 KO MEFs.

Amino-acid, glucose and serum starvation and stimulation of cells. For amino acids and/or glucose deprivation in MEFs, sub-confluent cells were rinsed twice and incubated in RPMI without amino acids, glucose or both, and supplemented with 10% dialysed FBS, as described[3]. Stimulation with glucose (5 mM) or amino acids (concentration as in RPMI) was performed for 10 min, unless otherwise indicated. For serum withdrawal, cells were rinsed twice in serum-free DMEM and incubated in serum-free DMEM for the indicated times; 100 nM was used for insulin stimulation. Aminoimidazole carboxamide ribonucleotide (AICAR, EMD Biosiences) was used at a final concentration of 2 mM. For cell survival experiments, cells were deprived of glucose or amino acids, and attached cells were counted 48 h later. For treatments with *in vivo* concentration of nutrients, MEFs were incubated with the following concentrations of amino acids (all in μM), reflecting the values found at birth (0 h) and after fasting for 1 h (1 h) in control mice: 0 h: D, 34; T, 446; S, 268; N, 180; E, 194; Q, 1221; P, 289; G, 382; V, 321; C, 26; M, 245; I, 122; L, 192; Y, 165; F, 189; W, 124; K, 1026; H, 74; R, 199; 1 h: D, 48; T, 172; S, 82; N, 57; E, 128; Q, 592; P, 183; G, 298; V, 154; C, 17; M, 164; I, 26; L, 37; Y, 71; F, 72; W, 92; K, 723; H, 30; R, 78. Similarly, the concentrations of glucose were as follows: 0 h: 45 mg dl^{-1}; 1 h: 12 mg dl^{-1}. Protein extracts were obtained as above.

Immunoblotting. Reagents were obtained from the following sources: anti phospho-T389 S6K1, phospho-S240/244 S6, phospho-T37/T46 4E-BP1, phosho-T308 Akt, phospho-S473 Akt, phospho-S757 ULK1, phospho-S9 GSK3-β, phospho-S641 glycogen synthase, phospho-S51-eIF2a, total Akt, S6K1, 4E-BP1, GSK3-β, glycogen synthase and eIF2a from Cell Signaling Technology; anti LC3B from Cell Signaling Technology and Novus Biologicals; anti β-actin (clone AC-15) from Sigma; anti p62 from America Research Products, Cell Signaling Technology and Enzo Life Sciences; anti PYGL from Santa Cruz. Cells were rinsed once with ice-cold PBS and lysed in ice-cold lysis buffer (50 mM HEPES (pH 7.4), 40 mM NaCl, 2 mM EDTA, 1.5 mM sodium orthovanadate, 50 mM NaF, 10 mM pyrophosphate, 10 mM glycerophosphate and 1% Triton X-100, and one tablet of EDTA-free complete protease inhibitors (Roche) per 25 ml). Cell lysates were cleared by centrifugation at 15,000*g* for 10 min. Protein extracts were denatured by the addition of sample buffer, boiled for 5 min, resolved by SDS–polyacrylamide gel electrophoresis and analysed by immunoblotting.

Immunofluorescence assays in cells. MEFs or HEK-293T cells were plated on fibronectin-coated 2 cm^2 glass coverslips at a density of 50,000–100,000 cells per coverslip. For overexpression of GFP–LC3B and TFEB–GFP, cells were transfected with nucleofection (Lonza) using 1 μg for 2×10^6 cells. The following day, cells were transferred to amino-acid- or glucose-free RPMI, starved for 60 min or starved for 50 min and re-stimulated for 10 min with amino acids or glucose, rinsed with cold PBS once and fixed for 15 min with 4% paraformaldehyde. Coverslips were permeabilized with 0.05% Triton X-100 in PBS and then incubated with primary antibodies in 5% normal donkey serum for 1 h, rinsed and incubated with Alexa Fluor-conjugated secondary antibodies (Invitrogen) diluted 1:400, for 45 min. Cells overexpressing GFP–LC3B and TFEB–GFP were fixed in 4% paraformaldehyde, rinsed and imaged. Coverslips were mounted on glass slides using Vectashield (Vector Laboratories) and imaged on a spinning disk confocal system (Perkin Elmer) equipped with 405, 488 and 561 nm laser lines, through a $\times$63 objective.

Co-immunoprecipitation assays. HEK-293T cells stably expressing Flag-tagged proteins were processed as described[5].

Quantitative PCR. Total RNA was extracted with RNeasy (Qiagen), retro-transcribed with Superscript III (Invitrogen) and used at 1:100 dilution in quantitative real-time PCR in an Applied Biosystems thermocycler. 36B4 and β-actin were for normalization. The following primers were used: *RagA* F, GAACCTGGTGCTGAACCTGT; *RagA* R, GATGGCTTCCAGACACGATT; *RagB* F, TTCGATTTCTGGGAAACCTG; *RagB* R, AGTTCACGGCTCTCCACATC; mTOR F, GGTGCTGACCGAAATGAGGG; mTOR (also known as *Mtor*) R, TCTTGCCCTTGTGTTCTGCA; *Raptor* (also known as *Rptor*) F, TGGCAGCCAAGGGCTCGGTA; *Raptor* R, GCAGCAGCTCGTGTGCCTCA; *Rictor* F, TCGCAACTCACCACAAGCGGG; *Rictor* R, TGCAAGCATCTGTGGCTGCCG; *Pepck1* F, CGATGACATCGCCTGGATGA; *Pepck1* R, TCTTGCCCTTGTGTTCTGCA; *G6pc* F, GAAGGCCAAGAGATGGTGTGA; *G6pc* R, TGCAGCTCTTGCGGTACATG; glucokinase F, GAGATGGATGTGGTGGCAAT; glucokinase R, ACCAGCTCCACATTCTGCAT; *36B4* (also known as *Rplp0*) F, TAAAGACTGGAGACAAGGTG; *36B4* R, GTGTACTCAGTCTCCACAGA; *Sqstm1* F, GAACTCGCTATAAGTGCAGTGT; *Sqstm1* R, AGAGAAGCTATCAGAGAGGTGG; *Vps11* F, GGAGCCTGGTCTTTGGAGA; *Vps11* R, GCTGTAGAGAACGTGGCAAGA; *Vps33a* F, TCTGTGCTCAGCAAGAAGGCGA; *Vps33a* R, GGACGCAAACTGCTTGATCTCC; *Vps8* F, GATGGACCATCTCCTGAAACAGG; *Vps8* R, AGCCTTCCTCTTGCTGACATCC; *Uvrag* F, GGAATAATGCCGGATCGTCTG; *Uvrag* R, CCTTTCCACCCCAAATCTTCAC; actin F, GGCACCACACCTTCTACAATG; actin R, GTGGTGGTGAAGCTGTAGCC.

Neonatal fasting and treatments. E17.5 and E18.5 pregnant females were injected with 2 mg progesterone (Sigma-Aldrich) to prevent early delivery. At E19.5, females were euthanized and fetuses immediately obtained by Caesarean section. Successfully resuscitated neonates were placed in a humidified chamber at 30 °C and fasted. Rapamycin (LC Laboratories) was administered intraperitoneally at a volume of 100 μl (1 mg ml^{-1} concentration) to pregnant females 4 h before Caesarean section, and neonates were injected subcutaneously immediately after Caesarean section. Glucose (30%) in PBS, or gluconeogenic amino acid mix (A, 500 mg ml^{-1}; N, 10 mg ml^{-1}; S, 6 mg ml^{-1}; D, E and P, 4 mg ml^{-1}; G, 2 mg ml^{-1}) in PBS, were injected subcutaneously every 3–6 h.

Electron microscopy. Tissues were obtained and immediately fixed in 2% glutaraldehyde in 0.1 M sodium cacodylate buffer pH 7.4 at room temperature. After post-fixation in 2% OsO_4, blocks were processed for embedding in Epon 812. Thin sections were obtained, stained with uranyl acetate and lead citrate, and examined by transmission electron microscopy in a JEOL EX 1200 electron microscope.

Measurement of glucose and amino-acid concentrations. Blood glucose was quantified with a glucometer and glucose test strips (Bayer Contour), with a lower detection limit of 10 mg dl^{-1}. For amino-acid quantification, plasma was analysed using the Waters MassTrak Amino Acid system. Pre-column derivatization of amino acids through molar excess of 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate was performed, converting both primary and secondary amino acids to stable chromophores. The derivatized amino acids were separated and detected using an Acquity UPLC system (Waters) and ultraviolet absorbance. Amino-acid concentrations in MEFs were quantified in the same manner after total extraction in boiling distilled H_2O.

Hepatic glycogen content measurement. Glycogen was measured in liver samples as described[23]. Briefly, glycogen was extracted from neonatal livers in 30% KOH saturated with Na_2SO_4, precipitated in 95% ethanol and re-suspended in double-distilled H_2O. After addition of phenol and H_2SO_4, absorbance at 490 nm was measured in triplicates.

Statistical analyses. For Kaplan–Meier survival curves, comparisons were made with the log-rank Mantel–Cox method. For quantitative PCR, measurements of glycaemia, plasma amino acids and glycogen content, non-parametric *t*-tests were performed. χ^2 tests were also performed for the effects of rapamycin on glycaemia.

doi:10.1038/nature11738

Visualization of splenic marginal zone B-cell shuttling and follicular B-cell egress

Tal I. Arnon[1], Robert M. Horton[1], Irina L. Grigorova[1]† & Jason G. Cyster[1]

The splenic marginal zone is a unique microenvironment where resident immune cells are exposed to the open blood circulation[1,2]. Even though it has an important role in responses against blood-borne antigens, lymphocyte migration in the marginal zone has not been intravitally visualized due to challenges associated with achieving adequate imaging depth in this abdominal organ. Here we develop a two-photon microscopy procedure to study marginal zone and follicular B-cell movement in the live mouse spleen. We show that marginal zone B cells are highly motile and exhibit long membrane extensions. Marginal zone B cells shuttle between the marginal zone and follicles with at least one-fifth of the cells exchanging between compartments per hour, a behaviour that explains their ability to deliver antigens rapidly from the open blood circulation to the secluded follicles. Follicular B cells also transit from follicles to the marginal zone, but unlike marginal zone B cells, they fail to undergo integrin-mediated adhesion, become caught in fluid flow and are carried into the red pulp. Follicular B-cell egress via the marginal zone is sphingosine-1-phosphate receptor-1 (S1PR1)-dependent. This study shows that marginal zone B cells migrate continually between marginal zone and follicles and establishes the marginal zone as a site of S1PR1-dependent B-cell exit from follicles. The results also show how adhesive differences of similar cells critically influence their behaviour in the same microenvironment.

Marginal zone B cells are a unique B-cell subset that have a pivotal role in mounting antibody responses against systemic pathogens[3,4]. Early studies of marginal zone B cells in rodents showed that they are non-recirculating and restricted to the spleen[5]. Marginal zone B cells were later found to have elevated integrin expression and to depend on integrins to be retained in the marginal zone[6]. These observations indicated that the cells were of limited motility. Yet, marginal zone B cells mediate the delivery of opsonized antigens from marginal zone to the follicle[7–9], and recent studies provided indirect evidence that marginal zone B cells continually exchange between the marginal zone and follicle[9,10]. However, this cellular behaviour has not been directly visualized. To permit real-time imaging of marginal zone B cells, we developed a way to label these cells. Follicular B cells can give rise to marginal zone B cells[4,11,12], and marginal zone B cells, but not follicular B cells, are self-renewing in the absence of input from less committed precursors[13]. We therefore asked whether follicular B cells could selectively reconstitute the marginal zone of CD19-deficient mice that have an empty marginal zone B-cell niche, but a normal follicle compartment[12,14,15]. Transfer of GFP$^+$ B cells into *Cd19* knockout mice for 8 weeks allowed substantial reconstitution of the marginal zone B-cell compartment (Fig. 1a). Moreover, typically ~90% of the transferred GFP$^+$ cells had a marginal zone B-cell phenotype (Fig. 1b and Supplementary Fig. 1). Like their normal counterparts[16], the reconstituted marginal zone B cells were poised to respond to antigen and lipopolysaccharide (Supplementary Fig. 1). To determine if the reconstituted marginal zone B cells were positioned correctly, we labelled blood-exposed cells by intravenous (i.v.) injection of a fluorescently conjugated antibody 5 min before tissue isolation[9,10]. This analysis, as well as immunofluorescence microscopy, indicated that 50–60% of the marginal zone B cells were in the marginal zone, whereas the remaining cells were located in the follicles (Fig. 1c, d), similar to their distribution in wild-type mice[9,10]. Moreover, the reconstituted mice were rescued in their ability to deposit an opsonized antigen on follicular dendritic cells (FDCs) over a 16-h period (Fig. 1e and Supplementary Fig. 1). Consistent with a direct role of the marginal zone B cells in the delivery of opsonized antigen to FDCs, reconstitution with *Cr2*$^{-/-}$ marginal zone B cells failed to restore antigen delivery (Supplementary Fig. 1).

For intravital two-photon laser-scanning microscopy (TPLSM), *Cd19* knockout mice reconstituted with a mixture of GFP$^+$ and non-labelled B cells were injected with red fluorescent phycoerythrin–immune complexes (PE–ICs) 2 h before imaging. Tissue section analysis established that PE–ICs were concentrated on SIGN-R1$^+$ (specific intracellular adhesion molecule-grabbing nonintegrin R1, also known as *Cd209b*) marginal zone macrophages in the first hours

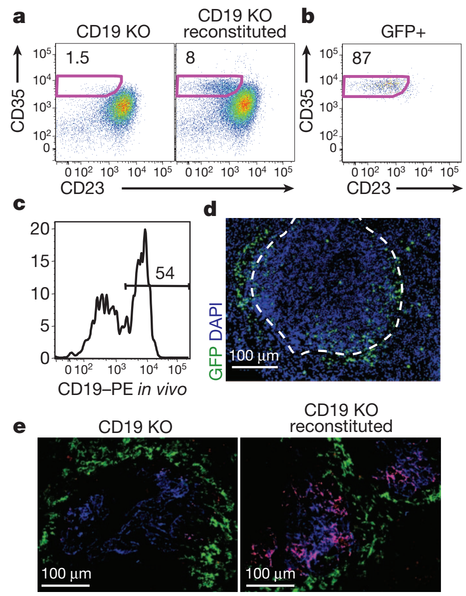


Figure 1 | Adoptive transfer system for GFP labelling marginal zone B cells. **a**, Frequency of CD35^{hi}CD23^{lo} marginal zone B cells among B220$^+$ cells in *Cd19*$^{-/-}$ mice before (left) or 8 weeks after (right) transfer of GFP$^+$ B cells. Numbers indicate percentage of cells in gate. **b**, Phenotype of CD19$^+$GFP$^+$ B cells from **a**. **c**, *In vivo* anti-CD19–PE labelling of marginal zone-phenotype (CD35^{hi}CD23^{lo}) B cells. Number indicates percentage of labelled cells. **d**, Spleen section from mouse reconstituted with a 2:1 mixture of non-transgenic and GFP$^+$ B cells, stained with anti-GFP (green) and 4′,6-diamidino-2-phenylindole (DAPI, blue). Location of marginal sinus is indicated by the dashed white line. **e**, Spleen sections from the indicated mice that had received PE–IC (red) 16 h earlier, stained for CD169 (green) to label metallophilic marginal zone macrophages and CD35 (blue) to label FDCs. Scale bar, 100 μm.

[1]Howard Hughes Medical Institute and Department of Microbiology and Immunology, University of California, San Francisco, California 94143, USA. †Current address: Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan 48109-5620, USA.

after injection, providing a means for locating this compartment (Fig. 2a). The spleen was surgically exposed, bathed in saline and stabilized by attachment to a platform placed over the mouse abdomen. Typically one or two white pulp cords per spleen passed sufficiently near the capsule to permit visualization (Supplementary Fig. 2). Marginal zone B cells were identified as being situated in the marginal zone or follicle based on whether their location overlapped with or was internal to the ring of PE–IC-labelled macrophages, respectively (Fig. 2b, c). Contours were drawn immediately internal to the PE–IC labelled cells in each z-plane and used to generate a three-dimensional surface (Fig. 2b) that approximated the position of the marginal sinus separating the marginal zone and follicle[1,2].

Marginal zone B cells within the marginal zone or follicle were migratory (Fig. 2c, d and Supplementary Videos 1 and 2). They travelled with similar velocities in both compartments, but cells within the marginal zone showed sharper turning angles and more deviation from movement in a straight direction in their migration paths (Fig. 2d), indicating a greater amount of confinement. Marginal zone B cells were larger than follicular B cells, as expected[16], and they exhibited a probing, dendritic morphology (Fig. 2e and Supplementary Videos 1 and 2) reminiscent of that seen for germinal centre B cells[17]. In some cases (~20%) the marginal zone B cells exhibited trailing cellular processes of remarkable length, occasionally exceeding 40 µm (Fig. 2c, e and Supplementary Video 2). Similar morphologies were observed for cells located in the marginal zone and follicle.

Treatment with FTY720 to disrupt S1PR1 function causes marginal zone B cells to leave the marginal zone and locate within the follicle[18]. Analysis of this repositioning at four time points using CD19–PE labelling of blood-exposed spleen cells suggested that it was complete within 30 min (Fig. 2f and Supplementary Fig. 3). A similar rate of marginal zone B-cell relocation was observed by TPLSM (Fig. 2g, Supplementary Fig. 3 and Supplementary Video 3), indicating that the same behaviour also occurred during the intravital imaging procedure.

To examine the rate of marginal zone B-cell movement between marginal zone and follicle in untreated mice, the tracks of cells that crossed between zones were manually annotated and counted (Fig. 3). Marginal zone B cells moving from the marginal zone across the boundary into the follicle were readily observed (Fig. 3a and Supplementary Video 4). Marginal zone B cells could also be seen migrating from the follicle to the marginal zone (Fig. 3b and Supplementary Video 5). Similar observations were made when the boundary between the follicle and marginal zone was defined using transferred follicular B cells rather than by PE–IC labelling (Supplementary Fig. 4), indicating that the migratory behaviour of marginal zone B cells was not a consequence of exogenous immune complex exposure. This analysis showed that at least 10% of the marginal zone cells that were tracked during a 30-min imaging session moved from the marginal zone to follicle, and a similar fraction of the marginal zone cells that started in the follicle moved to the marginal zone (Fig. 3c). On some occasions during passage across the boundary, the marginal zone B cells paused (Supplementary Video 4, example 1, and Supplementary Video 5) and sometimes they moved parallel to the surface before crossing (Supplementary Video 4, example 2). During the crossing event, some cells showed an obvious constriction of the cell body (Fig. 3a, Supplementary Fig. 4b lower panel, and Supplementary Video 4, example 2). We also observed several cells that remained tethered near the marginal zone–follicle interface by a

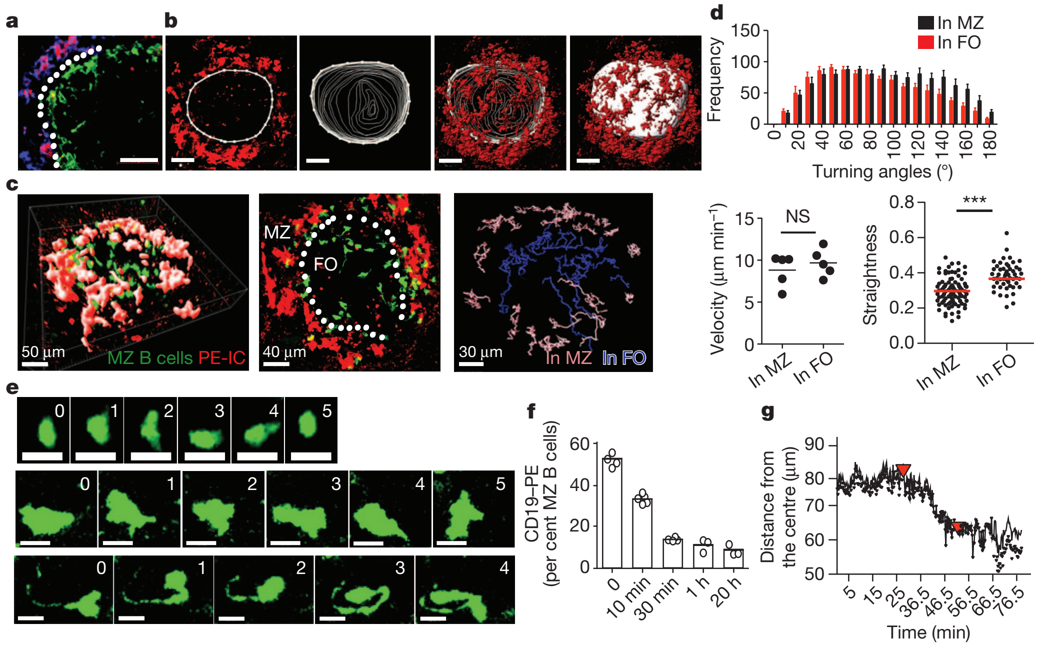


Figure 2 | Marginal zone B cells are migratory and exhibit long membrane processes. **a**, Spleen section showing PE–IC (red), SIGN-R1 (blue) and CD169 (green) distribution 2 h after PE–IC injection. White dotted line indicates the location of the marginal zone sinus. Scale bar, 50 µm. **b**, Generation of surface corresponding to the interface between the marginal zone and follicle. Left image shows an example of a contour drawn ~10 µm internal to the PE–ICs to represent the boundary in a single x–y slice (3 µm). Middle images show contours drawn for each slice in the 60 µm z-stack. Last image on the right shows the final surface with overlaid PE–IC stain. Scale bar, 50 µm. **c**, TPLSM of GFP^{+} marginal zone B cells in reconstituted *Cd19*$^{-/-}$ spleen. Left panel shows a 57 µm z-projection view. MZ, marginal zone. Middle panel shows a 30-µm slice from the centre of this region. FO, follicle. White dotted line indicates location of the marginal sinus. Right panel, representative classification of marginal zone B-cell tracks based on positioning with respect to surface. Pink, in marginal zone; blue, in follicle. **d**, Median velocity, distribution of turning angles and straightness of migration path of marginal zone B cells ($n = 5$ data sets from 3 mice). Straightness was calculated as a ratio of the total distance travelled divided by the displacement (difference between the initial and final position) for the first 5 min of each track. **e**, Time-lapse images of two marginal zone B cells (middle and bottom panels) compared with a follicular B cell (top panels). All cells are GFP^{+}. Scale bars, 10 µm. Time in min indicated on the panels. **f**, **g**, Kinetics of marginal zone B-cell displacement into the follicle following FTY720 treatment. **f**, Frequency of *in vivo* anti-CD19–PE labelled marginal zone B cells at the indicated time after FTY720 injection ($n = 6$ mice), detected by flow cytometry. Circles indicate individual data points and the bar indicates the mean. **g**, Average distance (micrometres) over time of GFP^{+} marginal zone B cells from the most central point of the follicle during TPLSM imaging. First red arrow at 25 min, time of FTY720 injection; second red arrow, time of anaesthetic reinjection. In **c**, **d** and **f**, bars or lines indicate means, error bars are s.e.m., ***$P < 0.0005$ and NS, not significant ($P > 0.05$) by unpaired Student's t-test.

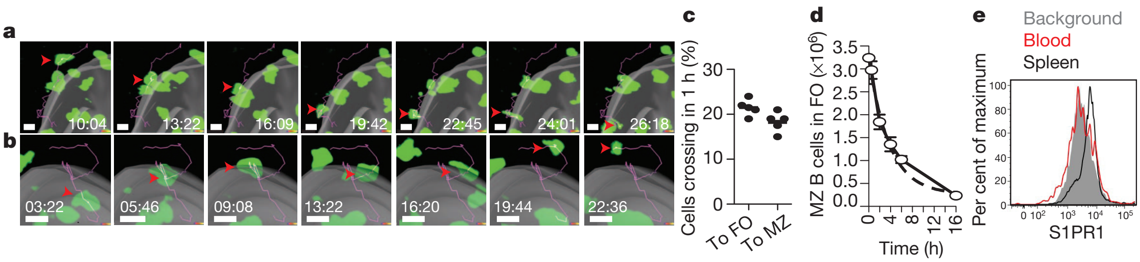


Figure 3 | Marginal zone B cells migrate bidirectionally between marginal zone and follicle. **a**, Marginal zone B-cell crossing from the marginal zone into the follicle. The grey surface represents the marginal zone–follicle interface. Time lapse is shown in min:s. Arrowheads show cell location at each time point and the pink line indicates the tracked path of the cell. Scale bar, 10 μm. **b**, Marginal zone B-cell migrating from the follicle to the marginal zone. Scale bar, 10 μm. **c**, Shuttling rate of marginal zone B cells (n = five data sets from three mice). **d**, Decay rate of marginal zone B cells from the follicle following integrin blockade. The number of marginal zone B cells remaining in the follicle was determined at each time point, showing a decay rate of 23% per hour (one representative experiment out of three shown). Error bars represent ± s.e.m. **e**, Flow cytometric analysis of S1PR1 on marginal zone B cells in spleen (black line) or after 1 h exposure to blood (red line).

membrane process while the cell body moved back and forth between zones (Supplementary Video 2, yellow dashed circle).

As an independent approach to estimate the rate of marginal zone B-cell movement from the follicles to marginal zone, in this case at the level of the whole spleen, we took advantage of the essential role of integrins in mediating marginal zone B-cell retention in the marginal zone. Treatment with integrin-blocking antibodies causes selective loss of cells from the marginal zone while not displacing cells that are situated within follicles[6]. Using this approach, loss of marginal zone B cells from follicles would occur over time as cells move from the follicle into the marginal zone. Mice were treated with integrin-neutralizing antibodies and the rate of marginal zone B-cell decay from follicles was determined by enumerating the *in vivo* CD19–PE unlabelled marginal zone B cells remaining in the spleen over time. The decay rate matched first order kinetics with a $t^{1/2}$ of around 2.5 h (Fig. 3d) consistent with the estimate from TPLSM of 20% exchange between the follicle and marginal zone per hour.

An exchange rate of 20% per hour indicates that some marginal zone B cells remain within the marginal zone for several hours. S1PR1 is required for marginal zone B cells to remain in the marginal zone[18] and when it is down-modulated by treatment with FTY720 the cells relocalize into the follicle within around 30 min (Fig. 2f, g). The level of S1PR1 on marginal zone B cells in the spleen was higher than on marginal zone B cells that had been transferred into blood for 1 h (Fig. 3e) and was more similar to cells exposed *in vitro* to low nM sphingosine-1-phosphate concentrations (Supplementary Fig. 5). Red blood cells, the main source of sphingosine-1-phosphate in blood, were detectable in the marginal zone but were sparse compared to their density in blood vessels and in the red pulp (Supplementary Fig. 5). A lower interstitial sphingosine-1-phosphate concentration in the marginal zone may cause a more gradual or less complete S1PR1 down-modulation than occurs on cells in circulatory blood, enabling a dwell time of several hours in the marginal zone.

We next examined the migration dynamics of follicular B cells. Intravital TPLSM of the spleen a day after intravenous transfer of follicular B cells showed a marked concentration of the cells within follicles (Fig. 4a). The follicular B cells migrated with a similar speed to marginal zone B cells and the two types of cell showed similar displacement over time while migrating within the follicle (Fig. 4b). As well as cells confined to follicles, transferred B cells could be visualized in the adjacent red pulp (Fig. 4c). In many cases the cells appeared to be moving in a straight path away from the follicle (Fig. 4c). B cells within the red pulp had a more rounded morphology than cells within the follicle and their axis ratio was approximately 30% reduced (Fig. 4d). In contrast to the active migratory behaviour of B cells within the follicle, many B cells within the red pulp failed to show evidence of active migration, but were instead intermittently stationary or fast moving (Fig. 4e and Supplementary Video 6). Occasionally during a fast-moving step the cell would disappear from view, possibly indicating that it had passed into a red pulp venule to be flushed from the spleen (Supplementary Video 6). Manual tracking of 550 red-pulp B cells showed that although many were stationary during the imaging period, threefold more cells appeared to move passively (fast and tangentially) versus actively (slow and meandering) (Fig. 4f).

The pathway by which follicular B cells exit from splenic follicles is not defined. One possible route these cells might take is by crossing the marginal zone sinuses, the path taken by marginal zone B cells. In agreement with this hypothesis, tracking follicular B-cell migration with respect to the surface generated using PE–IC-labelled marginal zone macrophages allowed the identification of cells migrating into the marginal zone (Fig. 4c, g and Supplementary Fig. 6a). Frequently upon entering this region, the follicular B cells underwent a rapid linear movement, perhaps a consequence of being caught in a region of flow (Fig. 4g, Supplementary Fig. 6a and Supplementary Video 7). The cells were usually retained moments later; some cells then appeared non-migratory during the rest of the imaging session (Supplementary Video 7, example 1), whereas others did continue to move (Supplementary Video 7, example 3). Axis ratio measurements of cells that crossed between zones showed that wild-type follicular B cells promptly became rounded after crossing into the marginal zone (Fig. 4h).

To test whether the striking difference in marginal zone and follicular B-cell behaviour within the marginal zone was a consequence of integrin-mediated adhesion, we examined the behaviour of GFP^+ marginal zone B cells in reconstituted *Cd19* knockout mice in the first hours after treatment with integrin-neutralizing antibodies. Under these conditions, marginal zone B cells in the marginal zone frequently ceased active migration, became rounded and then moved fast and tangentially in the direction of the red pulp (Fig. 4 i–l, Supplementary Fig. 6 and Supplementary Video 8). This change in behaviour was associated with a sharp increase in displacement over time (Fig. 4k, l). By contrast, marginal zone B cells within the follicle continued to migrate and they maintained their long membrane extensions (Supplementary Fig. 6), although their displacement over time was slightly reduced (Fig. 4l and data not shown), as observed for integrin-deficient cells in lymph nodes[19,20].

Although S1PR1 and sphingosine-1-phosphate have been argued to have a role in lymphocyte egress from the spleen[21], this conclusion has been based on indirect assessments, and it has also been suggested that the spleen is distinct from other lymphoid organs in not being sensitive to egress inhibition by FTY720 (refs 22, 23). This lack of clarity arises in part because both cell entry to and exit from the spleen occur via the blood and because the S1PR1-dependent egress step has not been visualized. To test whether S1PR1 was required for follicular B-cell movement from follicle to marginal zone, we co-transferred fluorescently labelled wild-type and S1PR1-deficient B cells into wild-type recipients and performed TPLSM (Fig. 4m). Plotting the tracks of all

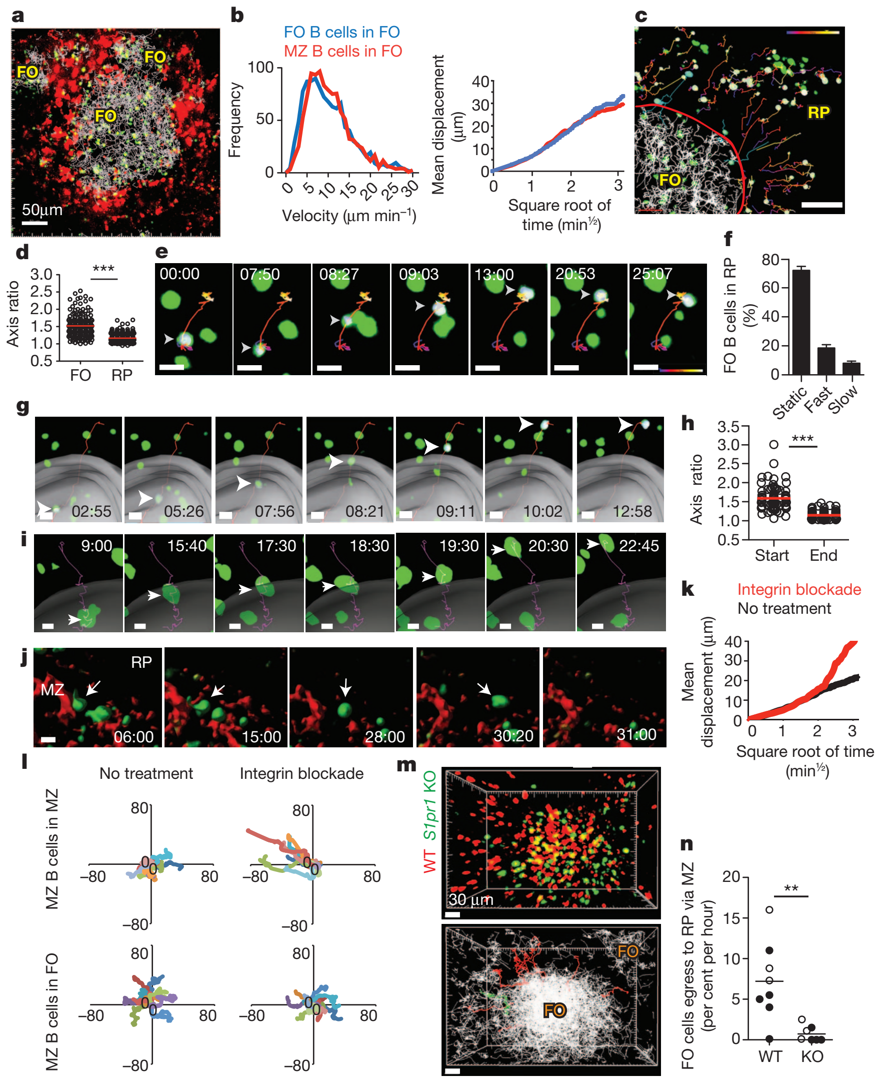


Figure 4 | Splenic follicular B-cell migration and S1PR1 requirement for exit. **a**, Fifty-four micrometre *z*-projection view showing transferred B cells (green-yellow) and their tracks (white lines) and PE–IC-labelled macrophages (red). Scale bar, 50 μm. **b**, Instantaneous velocities (left) and displacement versus square root of time (right) of follicle and marginal zone B cells in the follicle. Data from at least 5 or 6 experiments (3 or 4 mice). **c**, Forty-five micrometre *z*-projection view of follicular B cells (green) in follicle and red pulp (RP). Red line, marginal zone–follicle border; white lines, tracks of follicular B cells in the follicle; time-coded coloured lines (labelled as blue at the start and progressing to white at the end of the path), tracks of follicular B cells in red pulp; blue lines, follicular B cells in transition from the follicle to the marginal zone. Tracks of cells that were outside the follicle during the entire movie are highlighted with a white surface. Scale bar, 40 μm. **d**, Axis ratio of follicular B cells in the follicle and red pulp. **e**, Follicular B-cell movement in the red pulp. **f**, Percentage of cells in red pulp exhibiting stationary (static), rapid (fast) or migratory (slow) behaviour (5 experiments in 3 mice, $n = 550$ cells). **g**, Follicular B cell crossing from follicle to marginal zone. Grey surface, marginal zone–follicle interface. **h**, Axis ratio of follicular B cells migrating from follicle to marginal zone at the start and end of the track ($n = 24$ cells). **i–l**, Intravital TPLSM of marginal zone B cells following integrin blockade. **i**, Marginal zone B cell crossing from follicle to marginal zone. **j**, Marginal zone B-cell movement from marginal zone to red pulp. **k**, Displacement versus square root of time (right) of marginal zone B cells in the marginal zone before (black) and two hours after (red) integrin blockade. **l**, Superimposed 10-min tracks of randomly selected marginal zone B cells. The tracks show the path of each representative cell relative to the point of origin (00) in the *x*–*y* plane. Units are in micrometres. Plus or minus indicates migration in a particular direction. Data for **i–l** were from eight movies obtained from three mice. **m**, Upper image, 90 μm *z*-projection view of wild-type (red) and *S1pr1* knockout (KO) (green) follicular B cells in spleen. Lower image, automated tracks of transferred B cells (white). Tracks of wild-type cells (11 red lines) and *S1pr1* knockout cells (single green line) leaving the follicle are shown. **n**, Follicle egress rate of wild-type and *S1pr1* knockout B cells. Open circles, marginal zone–follicle interface determined based on PE–IC labelling; filled circles, interface determined based on follicular B cell tracks. In **e**, **g**, **i** and **j**, elapsed time is in min:s, arrowheads point to tracked cells and scale bar denotes 10 μm. In **d**, **f**, **h** and **n**, bars or lines represent the mean (error bars in **f** denote ± s.e.m.). In **d**, **h**, **n**, $^{**}P < 0.005$; $^{***}P < 0.0005$ by unpaired Student's *t*-test.

follicular B cells for a 30-min imaging session led to filling of the follicular area and the high-density region of tracks was used to generate a marginal sinus-approximating surface (Fig. 4m and Supplementary Fig. 7). S1PR1-deficient B cells moved within the follicle with similar velocities and turning angles to wild-type B cells (Supplementary Video 9 and data not shown). However, very few S1PR1-deficient B cells were observed exiting from the follicle into the marginal zone (Fig. 4m, n and Supplementary Video 9). In the rare cases where such movement was scored, no examples of cells undergoing the jump in movement were seen (Supplementary Video 9). A summary of these experiments showed that at least tenfold fewer S1PR1-deficient than wild-type B cells crossed out of the white pulp per hour (Fig. 4n).

The perpetual oscillatory movement of marginal zone B cells described here represents possibly the fastest rate of cellular exchange between tissue compartments so far observed and provides a mechanism for rapid delivery of opsonized blood-borne antigens into splenic follicles: marginal zone B cells capture antigens via complement receptors while travelling through the marginal zone and then transfer

them to follicular dendritic cells (FDCs) while migrating in the follicles. Remarkably, marginal zone B cells exhibit similar dynamics while resisting shear and migrating in an integrin-dependent fashion in the marginal zone and when moving in a largely integrin-independent fashion in the follicle. Precedent for a single cell type migrating with similar migration characteristics in an integrin-dependent and -independent manner is provided by findings using an *in vitro* system with dendritic cells[24]. The factors promoting formation of the long trailing cellular processes exhibited by many marginal zone B cells are unclear, but these membrane extensions may facilitate interactions with natural killer T cells[25] or with follicular B cells scanning for surface-displayed antigens[26]. Follicular B cells have lower integrin-adhesive activity than marginal zone B cells[6] and our data indicate that upon passage into the marginal zone they are unable to activate sufficient integrin activity to resist the local shear forces of blood flow and they become rounded and travel passively into the red pulp. As well as defining an S1PR1-dependent pathway of B-cell egress from the spleen, these results highlight how differences in the extent of adhesive interactions profoundly affect cell behaviour in the same microenvironment.

METHODS SUMMARY

Intravital imaging of marginal zone and follicular B cells in the spleen. B cells from Ub-GFP^{+} (4×10^6) or non-transgenic (8×10^6) mice were co-transferred to *Cd19*$^{-/-}$ recipient mice for 8–12 weeks. For imaging follicular B cells, purified B cells ($\sim 40 \times 10^6$) from B6 or *S1pr1*$^{fl/-}$*Mb1*$^{Cre/+}$ mice were labelled with fluorescent dye and transferred 24 h before imaging. To label marginal zone macrophages, mice were injected with PE–ICs 2 h before imaging. All imaging experiments were done intravitally using two-photon laser-scanning microscopy. To prepare for imaging, mice were anaesthetized[17], a skin incision was made below the costal margin and the spleen was gently exposed on its stalk. To immobilize the spleen, a spring-loaded platform[27] was placed over the mouse and screwed down. Saline was added to the contact area between the spleen and the coverslip. Tracks generated using Imaris Bitplane software were manually examined and verified.

Full Methods and any associated references are available in the online version of the paper.

Received 20 June; accepted 2 November 2012.
Published online 23 December 2012.

1. Schmidt, E. E., MacDonald, I. C. & Groom, A. C. Comparative aspects of splenic microcirculatory pathways in mammals: the region bordering the white pulp. *Scanning Microsc.* **7,** 613–628 (1993).
2. Mebius, R. E. & Kraal, G. Structure and function of the spleen. *Nature Rev. Immunol.* **5,** 606–616 (2005).
3. Martin, F. & Kearney, J. F. Marginal zone B cells. *Nature Rev. Immunol.* **2,** 323–335 (2002).
4. Pillai, S. & Cariappa, A. The follicular versus marginal zone B lymphocyte cell fate decision. *Nature Rev. Immunol.* **9,** 767–777 (2009).
5. MacLennan, I. C. M., Gray, D., Kumararatne, D. S. & Bazin, H. The lymphocytes of splenic marginal zones: a distinct B-cell lineage. *Immunol. Today* **3,** 305–307 (1982).
6. Lu, T. T. & Cyster, J. G. Integrin-mediated long-term B cell retention in the splenic marginal zone. *Science* **297,** 409–412 (2002).
7. Guinamard, R., Okigaki, M., Schlessinger, J. & Ravetch, J. V. Absence of marginal zone B cells in Pyk-2 deficient mice define their role in the humoral response. *Nature Immunol.* **1,** 31–36 (2000).
8. Ferguson, A. R., Youd, M. E. & Corley, R. B. Marginal zone B cells transport and deposit IgM-containing immune complexes onto follicular dendritic cells. *Int. Immunol.* **16,** 1411–1422 (2004).
9. Cinamon, G., Zachariah, M., Lam, O. & Cyster, J. G. Follicular shuttling of marginal zone B cells facilitates antigen transport. *Nature Immunol.* **9,** 54–62 (2008).
10. Arnon, T. I. *et al.* GRK2-dependent S1PR1 desensitization is required for lymphocytes to overcome their attraction to blood. *Science* **333,** 1898–1903 (2011).
11. Kumararatne, D. S. & MacLennan, I. C. Cells of the marginal zone of the spleen are lymphocytes derived from recirculating precursors. *Eur. J. Immunol.* **11,** 865–869 (1981).
12. You, Y., Zhao, H., Wang, Y. & Carter, R. H. Cutting edge: primary and secondary effects of CD19 deficiency on cells of the marginal zone. *J. Immunol.* **182,** 7343–7347 (2009).
13. Hao, Z. & Rajewsky, K. Homeostasis of peripheral B cells in the absence of B cell influx from the bone marrow. *J. Exp. Med.* **194,** 1151–1164 (2001).
14. Makowska, A., Faizunnessa, N. N., Anderson, P., Midtvedt, T. & Cardell, S. CD1high B cells: a population of mixed origin. *Eur. J. Immunol.* **29,** 3285–3294 (1999).
15. Martin, F. & Kearney, J. F. Positive selection from newly formed to marginal zone B cells depends on the rate of clonal production, CD19, and *btk*. *Immunity* **12,** 39–49 (2000).
16. Oliver, A. M., Martin, F., Gartland, G. L., Carter, R. H. & Kearney, J. F. Marginal zone B cells exhibit unique activation, proliferative and immunoglobulin secretory responses. *Eur. J. Immunol.* **27,** 2366–2374 (1997).
17. Allen, C. D., Okada, T., Tang, H. L. & Cyster, J. G. Imaging of germinal center selection events during affinity maturation. *Science* **315,** 528–531 (2007).
18. Cinamon, G. *et al.* Sphingosine 1-phosphate receptor 1 promotes B cell localization in the splenic marginal zone. *Nature Immunol.* **5,** 713–720 (2004).
19. Woolf, E. *et al.* Lymph node chemokines promote sustained T lymphocyte motility without triggering stable integrin adhesiveness in the absence of shear forces. *Nature Immunol.* **8,** 1076–1085 (2007).
20. Boscacci, R. T. *et al.* Comprehensive analysis of lymph node stroma-expressed Ig superfamily members reveals redundant and nonredundant roles for ICAM-1, ICAM-2, and VCAM-1 in lymphocyte homing. *Blood* **116,** 915–925 (2010).
21. Schwab, S. R. & Cyster, J. G. Finding a way out: lymphocyte egress from lymphoid organs. *Nature Immunol.* **8,** 1295–1301 (2007).
22. Rosen, H., Sanna, M. G., Cahalan, S. M. & Gonzalez-Cabrera, P. J. Tipping the gatekeeper: S1P regulation of endothelial barrier function. *Trends Immunol.* **28,** 102–107 (2007).
23. Morris, M. A. *et al.* Transient T cell accumulation in lymph nodes and sustained lymphopenia in mice treated with FTY720. *Eur. J. Immunol.* **35,** 3570–3580 (2005).
24. Schumann, K. *et al.* Immobilized chemokine fields and soluble chemokine gradients cooperatively shape migration patterns of dendritic cells. *Immunity* **32,** 703–713 (2010).
25. Barral, P., Sanchez-Nino, M. D., van Rooijen, N., Cerundolo, V. & Batista, F. D. The location of splenic NKT cells favours their rapid activation by blood-borne antigen. *EMBO J.* **31,** 2378–2390 (2012).
26. Suzuki, K., Grigorova, I., Phan, T. G., Kelly, L. & Cyster, J. G. Visualizing B cell capture of cognate antigen from follicular dendritic cells. *J. Exp. Med.* **206,** 1485–1493 (2009).
27. McDole, J. R. *et al.* Goblet cells deliver luminal antigen to CD103^{+} dendritic cells in the small intestine. *Nature* **483,** 345–349 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Miller for help with the microscope stage mount, T. Phan for advice regarding mouse surgery, J. An for technical assistance and O. M Bannard, J. R. Muppidi, M. Barnes and A. Reboldi for comments on the manuscript. T.I.A. was supported by a Jane Coffin Child's fellowship and J.G.C. is an Investigator of the Howard Hughes Medical Institute. This work was supported in part by National Institutes of Health grant AI74847.

Author Contributions T.I.A. and J.G.C. conceived and designed the experiments. T.I.A. performed the experiments. R.M.H. helped with some of the quantitative image analysis. I.L.G. helped with early aspects of the spleen surgery procedure. T.I.A. and J.G.C. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.G.C. (Jason.Cyster@ucsf.edu).

METHODS

Mice. 6–12-week-old C57BL/6 (B6) mice were purchased from the National Cancer Institute. *S1pr1*$^{fl/-}$ mice (R. Proia) were crossed with *Mb1*$^{Cre/+}$ mice (M. Reth) to generate *S1pr1*$^{fl/-}$*Mb1*$^{Cre/+}$. B6 mice expressing enhanced green fluorescent protein (GFP; 004353; Tg(UBC-GFP)30Scha/J) were from Jackson Laboratories. *Cd19*$^{-/-}$ mice were in a B6 background and were generated by intercrossing *Cd19*$^{Cre/+}$ mice[28] to obtain *Cd19*$^{Cre/Cre}$ mice. *Cr2*$^{-/-}$ mice[29] were in a B6 background. Animals were housed in a specific pathogen-free facility and all experiments were in accordance with protocols approved by the University of California San Francisco Institutional Animal Care and Use Committee.

B-cell isolation, adoptive transfer and selective reconstitution of labelled marginal zone B cells. B cells were isolated with an AutoMACS (Miltenyi Biotech) using MACS microbeads and antibodies to CD11c and CD43 (Miltenyi Biotech). Purity was typically over 95%. To label marginal zone B cells selectively, 4×10^6 B cells from a Ub-GFP$^+$ donor and 8×10^6 B cells from non-transgenic B6 mice were mixed and transferred to *Cd19*$^{-/-}$ recipient mice for 8–12 weeks. The phenotype of the cells remained constant across this ~1 month analysis window. For imaging of follicular B cells, purified B cells from B6 or *S1pr1*$^{fl/-}$*Mb1*$^{Cre/+}$ were labelled with 10 μM 5-(and-6)-(((4 chloromethyl)benzoyl)amino)-tetramethylrhodamine (CMTMR, Invitrogen) or carboxyfluorescein diacetate succinimidyl ester (CFDA-SE, Invitrogen). In some experiments B cells were purified from GFP$^+$ donors and used. Cells (~40×10^6) were transferred 24 h before imaging.

Flow cytometry, *in vivo* labelling and immunofluorescence of cryostat sections. Marginal zone B cells were gated as B220$^+$ or CD19$^+$CD23loCD21^{hi}CD35^{hi} and follicular B cells as B220$^+$ or CD19$^+$CD23^{hi}CD21^{int}CD35^{int}. CD19, B220, CD23, IgM, IgD, CD1d, CD86, CD38 and CD44 antibodies were from BioLegend. CD35 antibody was from BD Biosciences. *In vivo* labelling of marginal zone B cells was as described[30,31]. Immunofluorescence analysis of spleen sections was as described[32]. Marginal metallophilic macrophages were detected with CD169-FITC antibody (AbD Serotec) and marginal zone macrophages with SIGN-R1 antibody (eBioscience) followed by anti-Armenian hamster Cy5 (Jackson Immunoresearch). Follicular dendritic cells (FDCs) were detected with anti-complement receptor-1 (CD35) antibody (BioLegend). For spleen sections containing GFP$^+$ cells, spleens were fixed in 4% paraformaldehyde (Sigma-Aldrich) and stained with GFP antibody (Invitrogen) and DAPI (Sigma-Aldrich).

PE–IC deposition on follicular dendritic cells. PE–ICs were induced *in vivo* by passively immunizing mice with 2 mg polyclonal rabbit IgG anti-PE (200-4199; Rockland) followed 2 h later by injection of 20 μg PE (P-801, Invitrogen Molecular Probes) as described[33]. After 16 h, spleens were collected and frozen in optimum cutting temperature (OCT) compound (VWR International) for immunofluorescence analysis.

Marginal zone macrophage labelling *in vivo* for intravital TPLSM. We attempted to visualize the boundary between the marginal zone and follicle by intravital labelling of the marginal sinus-lining cells with MAdCAM1-specific antibody Meca379 (ref. 34). However, the intensity of marginal sinus labelling was insufficient to allow detection in the intact spleen by TPLSM (not shown). Instead, we found that when mice were injected with Meca379 antibody (Bio X Cell) that had been biotinylated using biotin X-NHS (EMD Chemical) and mixed with streptavidin-PE (Bio-Rad Laboratories) in a 4:1 ratio, and were examined 2–6 h later, there was prominent labelling of marginal zone macrophages (which lack MAdCAM1), whereas the CD169$^+$ marginal metallophillic macrophages that are mostly positioned on the follicular side of the marginal sinus[2,35] were unlabelled. Because the labelling achieved by the biotinylated antibody–PE-streptavidin complexes was similar to that by immune complexes[8,33], for simplicity we refer to this method as labelling with PE–ICs. Mice were injected with the premade PE–ICs (100 μg) 2 h before the beginning of imaging. Although this approach allowed detection of heavily PE–IC-decorated marginal zone macrophages, it was of insufficient sensitivity to permit TPLSM detection of PE–IC bearing B cells in the spleen (not shown).

Integrin blockade assays. Anti-αL (clone M17/4, rat IgG2a) hybridoma was from American Type Culture Collection, and the anti-α4 (clone PS/2, rat IgG2b) hybridoma was provided by D. Erle. Antibodies were injected intravenously at 100 μg per mouse for the indicated amounts of time followed by 5-min CD19–PE *in vivo* labelling as described before. In TPLSM experiments, the antibodies were injected 15 min to 3 h before image acquisition.

Surface expression of S1PR1 on marginal zone B cells in the spleen and blood. Freshly collected splenocytes from CD45.1$^+$ donors (10×10^6 cells) were transferred into CD45.2$^+$ recipients. S1PR1 expression on donor marginal zone B cells in the spleen was determined before transfer and in blood collected from recipients 1 h after transfer. Expression of S1PR1 was detected by flow cytometry using a rat monoclonal antibody (R&D Systems) followed by donkey anti-rat IgG biotin (Jackson ImmunoResearch Laboratories) and streptavidin-APC (Invitrogen) as described[36,37]. Background S1PR1 stain refers to stained samples from mice that were intravenously injected with the S1PR1-modulating drug FTY720 (1 mg kg^{-1}) 24 h before[38].

Intravital two-photon laser-scanning microscopy (TPLSM) of spleen. Mice were anaesthetized by intraperitoneal injection of 10 ml kg^{-1} saline containing xylazine (1 mg ml^{-1}) and ketamine (5 mg ml^{-1}). Maintenance doses of intramuscular injections of 4 ml kg^{-1} of xylazine (1 mg ml^{-1}) and ketamine (5 mg ml^{-1}) were given approximately every 30 min. To expose the spleen, a skin incision was made below the costal margin in the left flank overlying the spleen and extended inferomedially. An ~1 cm window was then made in the peritoneal cavity and the spleen was gently mobilized on its stalk with forceps and exteriorized without stretching or damaging the vessels in the hilum and the gastrosplenic ligament. After the spleen was exposed, the skin incision was partially closed with tissue glue (Vetbond) and the spleen was bathed in warm saline. A spring-loaded platform[27] was placed over the mouse and screwed down until the cover glass made contact with the spleen capsule. The spleen was kept almost sealed against the mouse body using the platform and attached coverslip, and the area around the spleen and in contact with the glass was kept filled with saline. The mouse was placed on a Biotherm stage warmer at 37 °C (Biogenics) for the duration of the imaging. The temperature at the interface between the spleen and glass coverslip during and at the end of several imaging sessions was measured using a dual-temperature controller (TC-344B, Warner Instruments) equipped with a CC-28 cable containing a bead terminator and was found to remain between 36–37 °C. Images were acquired with ZEN2009 (Carl Zeiss) using a 7MP two-photon microscope (Carl Zeiss) equipped with a Chameleon laser (Coherent). For video acquisition, a series of planes of 3 μm z-spacing spanning a depth of 50–150 μm were collected every 15–30 s. Excitation wavelengths were 870–890 nm. Emission filters were 500–550 nm for CFDA-SE and GFP, and 570–640 nm for PE and CMTMR. The full longitudinal extent of the spleen was surveyed in each animal at depths of ~50–100 μm and typically one or two white pulp cords were identified that passed sufficiently near the capsule to permit imaging of cells in the marginal zone and follicle. Videos were made and analysed with Imaris 7.4.2 ×64 (Bitplane). To track cells, surfaces seed points were created and tracked over time. Tracks were manually examined and verified. Data from cells that could be tracked for at least 15 min were used for analysis. The velocities, turning angles, and displacement of cells between each imaging frame were analysed using Imaris (Bitplane AG), MATLAB (MathWorks), and MetaMorph software. In Fig. 2d, graphs compare tracks that remained in the marginal zone or the follicle during 25–30 min of video acquisition. Marginal zone B cells that showed 'tethered oscillation' at the boundary (Supplementary Video 2) were not enumerated as crossing from one zone to the other as they did not travel a minimum of 10 μm into the opposite compartment. Statistical analysis was performed using Prism software (GraphPad Software). Annotation and final compilation of videos were performed with After Effects 7.0 software (Adobe). Video files were converted to MPEG format with AVI–MPEG Converter for Windows 1.5 (FlyDragon Software).

Analysis of marginal zone B cells distribution before and after FTY720-treatment. Marginal zone B cells were imaged intravitally for 25 min and then injected intravitally with 25 μg FTY720. Five minutes after injection, imaging of the same region was resumed for an additional 50 min. Marginal zone B cells were identified and their positions were determined using automatic segmentation in Imaris software (BitPlane AG). The coordinates of each cell were exported to a text file as comma separated values, and loaded into the R programming environment for analysis. For each time frame, the centre point of the population of cells was determined by taking the mean of the cell positions in x, y and z dimensions. The distance of each cell position to its time frame centre point was calculated, and the mean of these distances for each time frame were transferred to Microsoft Excel. The frame numbers from the movie were adjusted to reflect overall time during the experiment.

Axis ratio calculation. The long and short axis of the cells were measured in a single z plane via the line segment tool in Imaris software. Cells shape index was then calculated as the ratio of the longer axis to the shorter axis. For axis ratio measurements of cells migrating from the follicle to the marginal zone, each data point reflects the mean axis ratio of a single cell measured in the first (start) or last (end) three frames of the track.

Follicular B-cell egress. Wild-type (CMTMR-labelled, red) and *S1pr1* knockout (CFSE-labelled, green) B cells were co-transferred into a wild type recipient 24 h before intravital TPLSM. Both B-cell types were tracked, and cells that travelled a minimum of 10 μm into the marginal zone were scored. Each point corresponds to a single 30–60 min video, open circles to experiments where the marginal zone–follicle interface was determined based on PE–IC labelling (as in Fig. 2) and filled circles to experiments where the interface was determined based on follicular B-cell tracks.

28. Rickert, R. C., Rajewsky, K. & Roes, J. Impairment of T-cell-dependent B-cell responses and B-1 cell development in CD19-deficient mice. *Nature* **376,** 352–355 (1995).

29. Molina, H. *et al.* Markedly impaired humoral immune response in mice deficient in complement receptors 1 and 2. *Proc. Natl Acad. Sci. USA* **93,** 3357–3361 (1996).
30. Cinamon, G., Zachariah, M. A., Lam, O. M., Foss, F. W. Jr & Cyster, J. G. Follicular shuttling of marginal zone B cells facilitates antigen transport. *Nature Immunol.* **9,** 54–62 (2008).
31. Pereira, J. P., An, J., Xu, Y., Huang, Y. & Cyster, J. G. Cannabinoid receptor 2 mediates the retention of immature B cells in bone marrow sinusoids. *Nature Immunol.* **10,** 403–411 (2009).
32. Hargreaves, D. C. *et al.* A coordinated change in chemokine responsiveness guides plasma cell movements. *J. Exp. Med.* **194,** 45–56 (2001).
33. Phan, T. G., Grigorova, I., Okada, T. & Cyster, J. G. Subcapsular encounter and complement-dependent transport of immune complexes by lymph node B cells. *Nature Immunol.* **8,** 992–1000 (2007).
34. Kraal, G., Schornagel, K., Streeter, P. R., Holzmann, B. & Butcher, E. C. Expression of the mucosal vascular addressin, MAdCAM-1, on sinus-lining cells in the spleen. *Am. J. Pathol.* **147,** 763–771 (1995).
35. Kraal, G. & Mebius, R. New insights into the cell biology of the marginal zone of the spleen. *Int. Rev. Cytol.* **250,** 175–215 (2006).
36. Green, J. A. & Cyster, J. G. S1PR2 links germinal center confinement and growth regulation. *Immunol. Rev.* **247,** 36–51 (2012).
37. Arnon, T. I. *et al.* GRK2-dependent S1PR1 desensitization is required for lymphocytes to overcome their attraction to blood. *Science* **333,** 1898–1903 (2011).
38. Lo, C. G., Xu, Y., Proia, R. L. & Cyster, J. G. Cyclical modulation of sphingosine-1-phosphate receptor 1 surface expression during lymphocyte recirculation and relationship to lymphoid organ transit. *J. Exp. Med.* **201,** 291–301 (2005).

doi:10.1038/nature11776

Reciprocal regulation of p53 and malic enzymes modulates metabolism and senescence

Peng Jiang[1]*, Wenjing Du[1]*, Anthony Mancuso[1], Kathryn E. Wellen[1] & Xiaolu Yang[1]

Cellular senescence both protects multicellular organisms from cancer and contributes to their ageing[1]. The pre-eminent tumour suppressor p53 has an important role in the induction and maintenance of senescence, but how it carries out this function remains poorly understood[1–3]. In addition, although increasing evidence supports the idea that metabolic changes underlie many cell-fate decisions and p53-mediated tumour suppression, few connections between metabolic enzymes and senescence have been established. Here we describe a new mechanism by which p53 links these functions. We show that p53 represses the expression of the tricarboxylic-acid-cycle-associated malic enzymes ME1 and ME2 in human and mouse cells. Both malic enzymes are important for NADPH production, lipogenesis and glutamine metabolism, but ME2 has a more profound effect. Through the inhibition of malic enzymes, p53 regulates cell metabolism and proliferation. Downregulation of ME1 and ME2 reciprocally activates p53 through distinct MDM2- and AMP-activated protein kinase-mediated mechanisms in a feed-forward manner, bolstering this pathway and enhancing p53 activation. Downregulation of ME1 and ME2 also modulates the outcome of p53 activation, leading to strong induction of senescence, but not apoptosis, whereas enforced expression of either malic enzyme suppresses senescence. Our findings define physiological functions of malic enzymes, demonstrate a positive-feedback mechanism that sustains p53 activation, and reveal a connection between metabolism and senescence mediated by p53.

We previously found that p53 inhibits the important NADPH producer glucose-6-phosphate dehydrogenase[4]. As this did not fully explain the effect of p53 on NADPH, we investigated whether p53 controls the expression of malic enzymes, which catalyse the oxidative decarboxylation of malate to generate pyruvate and either NADPH or NADH[5,6] (Supplementary Fig. 1). In mammalian cells, three malic enzyme isoforms have been identified: a cytosolic $NADP^+$-dependent isoform (ME1), a mitochondrial $NAD(P)^+$-dependent isoform (ME2) and a mitochondrial $NADP^+$-dependent isoform (ME3), of which ME1 and ME2 are the main isoforms (Supplementary Fig. 2a)[7]. By recycling the tricarboxylic acid (TCA) cycle intermediate malate into the common TCA cycle carbon source pyruvate, malic enzymes may have a regulatory role in matching TCA flux to cellular demand for energy, reducing equivalents and biosynthetic precursors (Supplementary Fig. 1).

We knocked down *TP53* in human osteosarcoma U2OS cells and normal diploid fibroblast IMR90 cells using short hairpin RNA (shRNA). This led to a significant increase in messenger RNA levels of *ME1* and *ME2* (Fig. 1a, b and Supplementary Fig. 2b), accompanied by increased protein levels and total enzymatic activity of ME1 and ME2 (Fig. 1a, c and Supplementary Fig. 2c). Likewise, expression of ME1 and ME2 were substantially higher in *Trp53* knockout ($Trp53^{-/-}$) compared to p53-wild-type ($Trp53^{+/+}$) mouse embryonic fibroblasts (MEFs) (Fig. 1d). The normally short-lived p53 protein is stabilized by DNA-damage signals. Cells treated with the genotoxic agents etoposide and doxorubicin showed both time- and concentration-dependent reductions in the expression of ME1 and ME2 (Fig. 1e and Supplementary Fig. 2b, d–g). When *TP53* was knocked down, the expression of ME1 and ME2 no longer responded to DNA damage (Fig. 1e). These results indicate that the expression of ME1 and ME2 is controlled by p53, both at basal levels and when p53 is stabilized by DNA-damage signals.

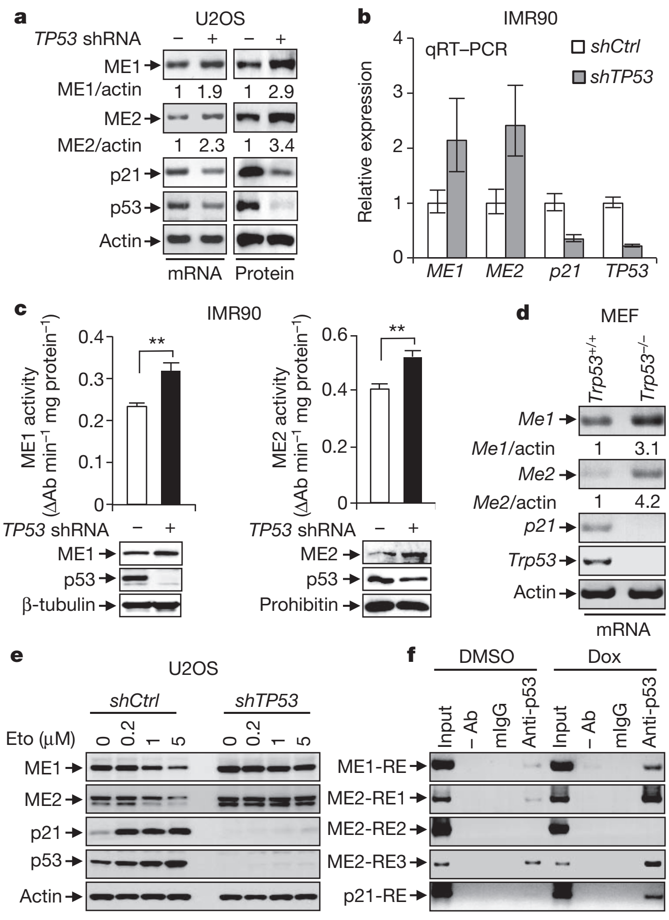


Figure 1 | p53 represses the expression of malic enzymes. a, Malic enzyme mRNA and protein expression in U2OS cells stably expressing *TP53* shRNA or control (Ctrl) shRNA. Relative malic enzyme/actin ratios are given. **b**, **c**, mRNA expression (**b**), total activity and protein levels (**c**) of malic enzymes in *TP53*-depleted and control IMR90 cells. Data shown are mean ± s.d. ($n = 3$). **d**, Malic enzyme gene expression in $Trp53^{+/+}$ and $Trp53^{-/-}$ MEFs. **e**, *TP53*-depleted and control U2OS cells were treated with increasing amounts of etoposide (Eto) and assayed for malic enzyme expression. **f**, $TP53^{+/+}$ HCT116 cells treated with or without doxorubicin (Dox; 1 μg ml^{-1}) were subjected to chromatin immunoprecipitation assay with anti-p53 DO-1 antibody, control mouse IgG (mIgG), or no antibody (− Ab). DMSO, dimethylsulphoxide; qRT–PCR, quantitative reverse-transcriptase PCR. $**P < 0.01$.

[1]Department of Cancer Biology and Abramson Family Cancer Research Institute, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

By analysing the malic enzyme gene sequences, we identified a putative p53 response element[8] in the first intron of the *ME1* gene (ME1-RE) and three putative response elements in the first intron of the *ME2* gene (ME2-RE1, ME2-RE2 and ME2-RE3) (Supplementary Fig. 3a). Chromatin immunoprecipitation assays in HCT116 cells revealed that p53 bound to the genomic regions of ME1-RE, ME2-RE1 and ME2-RE3, but not ME2-RE2. This binding was further enhanced when p53 was stabilized by treatment with doxorubicin (Fig. 1f). In addition, p53 repressed the expression of a luciferase gene driven by the genomic fragment containing ME1-RE, ME2-RE1 or ME2-RE3, but not ME2-RE2 (Supplementary Fig. 3b). p53-mediated repression of certain target genes involves histone deacetylases[9]. Treatment with trichostatin A, an inhibitor of histone deacetylases, abrogated p53-mediated repression of *ME1* and *ME2* genes (Supplementary Fig. 2g).

TP53 deficiency also led to a strong increase in the *ME3* transcript (Supplementary Fig. 4a). A putative p53 response element is present in the first intron of the *ME3* gene (ME3-RE) (Supplementary Fig. 4b). p53 bound to the genomic region of ME3-RE in cells (Supplementary Fig. 4c) and reduced the expression of a luciferase reporter driven by this response element (Supplementary Fig. 4d). Given the low abundance of ME3 expression in cell lines that have been tested (Supplementary Fig. 2a)[7], we focused on ME1 and ME2 in subsequent analyses.

Although ME1 and ME2 have been extensively characterized *in vitro*, there is a paucity of information on their functions in cells. Silencing *ME1* and *ME2*—in particular *ME2*—with short interfering RNA (siRNA) reduced cellular NADPH levels in IMR90 and U2OS cells (Fig. 2a and Supplementary Fig. 5a). This effect was also observed when a separate set of malic enzyme siRNAs, as well as malic enzyme shRNAs, were used (Supplementary Fig. 5b, c). By contrast, forced expression of ME1 or ME2—in particular ME2—or the addition of a malic enzyme substrate (dimethyl L-malate) increased cellular NADPH levels (Fig. 2b lanes 1–3, and Supplementary Fig. 5d). To determine whether the effect of malic enzymes is due to their enzymatic activity, we generated two ME1 mutations ($ME1^{mut1}$ and $ME1^{mut2}$) and three ME2 mutations ($ME2^{mut1}$, $ME2^{mut2}$ and $ME2^{mut3}$), each of which exhibited little or no enzymatic activity *in vitro* as well as *in vivo* (Supplementary Fig. 6). None of these mutants were able to increase cellular NADPH levels (Fig. 2b, lanes 4–10). Thus, both ME1 and ME2 are required for maintaining cellular NADPH levels through their enzymatic activity, with ME2 having a more profound effect. As previously observed[4], knockdown of *TP53* led to a significant increase in NADPH levels. This increase was partially reversed through the silencing of *ME1* and near-completely reversed through the silencing of *ME2* (Fig. 2a and Supplementary Fig. 5a, b). These results indicate that p53 regulates NADPH metabolism through the suppression of both malic enzymes, particularly ME2.

As NADPH provides reducing equivalents for reductive biosynthesis, we examined the role of malic enzymes in lipid production. MEF cells and murine-derived 3T3-L1 pre-adipocytes were cultured with a cocktail that stimulated their differentiation into adipocytes[4,10]. Triglycerides and total lipid levels in these cells declined when *Me1* or *Me2*—particularly *Me2*—was depleted in these cells (Fig. 2c and Supplementary Fig. 7). By contrast, overexpression of both enzymes, particularly ME2, but none of the ME1 or ME2 mutants, increased lipid abundance (Fig. 2d). Concordant with previously published data[4], we observed a marked increase in lipid levels in *Trp53*-deficient cells compared to *Trp53*-proficient cells. *Me1* knockdown partially reversed this increase, whereas *Me2* knockdown prevented it entirely, correlating with its greater influence on cellular NADPH levels (Fig. 2c and Supplementary Fig. 7). These results indicate that the enhanced lipid accumulation in p53-deficient cells is dependent on the malic enzymes, especially ME2.

Silencing of *ME2*, as well as of *ME1*, did not significantly alter NADH levels or the NAD^+/NADH ratio in IMR90 cells (Supplementary

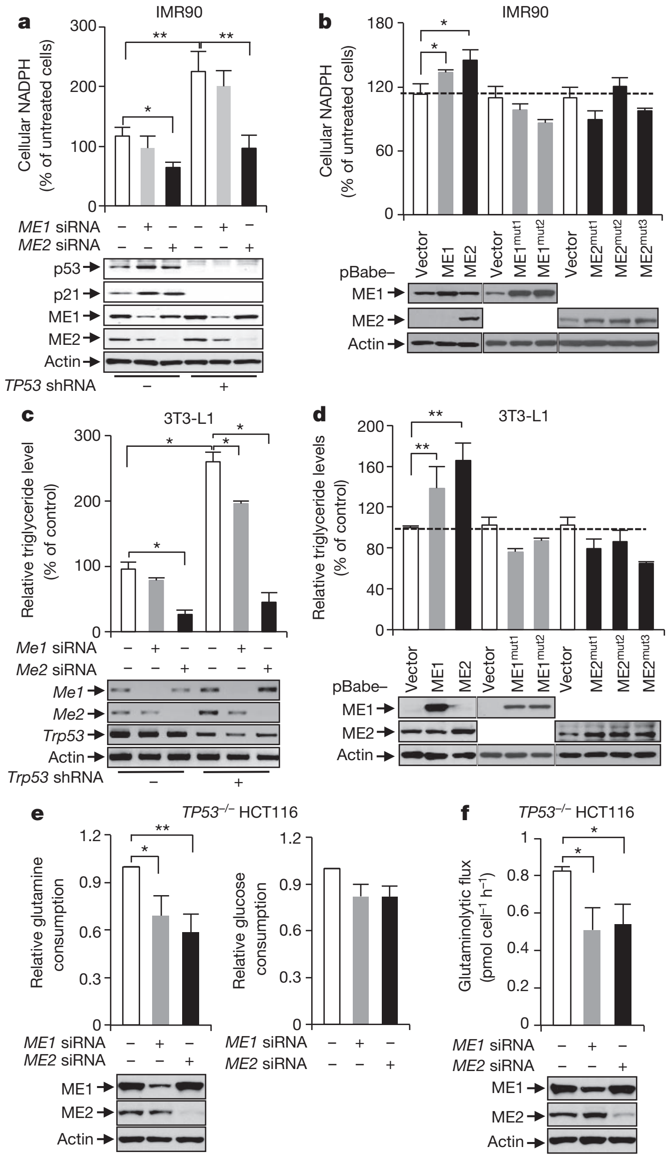


Figure 2 | ME1 and ME2 influence NADPH production, lipid production and glutaminolysis. **a**, **b**, NADPH levels in *TP53*-depleted and control IMR90 cells transfected with control, *ME1* or *ME2* siRNA (**a**), or in IMR90 cells stably overexpressing wild-type malic enzymes, mutant malic enzymes or vector control (**b**). Protein expression is shown below. Stable protein expression was achieved using the pBabe retroviral expression system. **c**, **d**, Triglyceride contents in *Trp53*-depleted and control 3T3-L1 cells transfected with control, *Me1* or *Me2* siRNA (**c**), or 3T3-L1 cells stably expressing wild-type malic enzymes, mutant malic enzymes or vector control (**d**). mRNA (**c**) and protein (**d**) expression is shown below. **e**, **f**, Effect of *ME1* and *ME2* knockdown in $TP53^{-/-}$ HCT116 cells on glucose and glutamine consumption (**e**) and glutaminolytic flux (**f**). Protein expression is shown below. All error bars represent mean ± s.d. ($n = 3$). $*P < 0.05$; $**P < 0.01$.

Fig. 8a, b), despite the fact that ME2 is characterized as either $NADP^+$ or NAD^+ dependent[5,6]. NADH is the main electron donor for the electron transport chain that drives ATP production. Silencing of either malic enzyme did not significantly alter the abundance of cellular ATP or ADP in IMR90 cells (Supplementary Fig. 8c). In U2OS cells silencing of *ME1*, but not *ME2*, reduced NADH levels and increased the NAD^+/NADH ratio (Supplementary Fig. 8d, e). These results are consistent with a cell-type-specific role of ME1 and a minimal role of ME2 in maintaining cellular NADH and ATP levels.

Next we investigated the role of ME1 and ME2 in the metabolism of glucose and glutamine. In $TP53^{-/-}$ HCT116 cells silencing of either malic enzyme, but especially *ME2*, strongly reduced glutamine

consumption (Fig. 2e), whereas silencing of either malic enzyme had a moderate effect on glucose consumption. We extended this analysis by evaluating the rate of glutaminolysis. Depletion of either *ME1* or *ME2* noticeably slowed down glutaminolytic flux (Fig. 2f). These results indicate that both ME1 and ME2 have a key role in glutamine metabolism but a relatively minor role in glucose metabolism.

p53 is critical for the induction and maintenance of senescence[1–3]. We noticed that in IMR90 cells, a well-established senescence model, silencing of each malic enzyme by either siRNA or shRNA caused a profound increase in cells expressing senescence-associated β-galactosidase, stopping growth (Fig. 3a, b and Supplementary Fig. 9a–f). The induction of senescence in malic-enzyme-knockdown cells was also indicated by the marked accumulation the promyelocytic leukaemia protein nuclear bodies[11,12] (Fig. 3c and Supplementary Fig. 9g). Notably, even a moderate reduction (20–30%) in either *ME1* or *ME2* strongly elicited senescence (Supplementary Fig. 10a). Malic-enzyme-loss-induced senescence also occurred in U2OS and *TP53*$^{+/+}$ HCT116 tumour cell lines (Supplementary Fig. 10b, c). In *TP53*-deficient primary and tumour cell lines senescence decreased markedly and malic enzyme depletion lost its ability

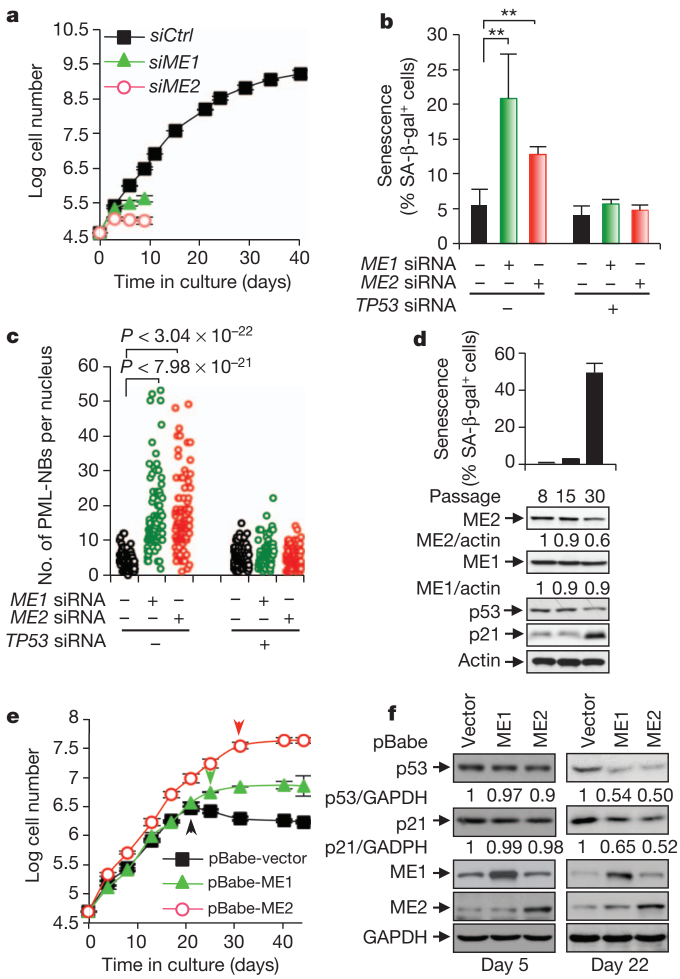


Figure 3 | A role for malic enzymes in suppressing p53-mediated senescence. **a**, The replicative lifespan of IMR90 cells transfected with control, *ME1* or *ME2* siRNA. **b**, **c**, IMR90 cells transfected with *ME1*, *ME2*, *TP53* or control siRNA as indicated. Percentages of senescence-associated β-galactosidase (SA-β-gal)-positive cells (**b**) and numbers of promyelocytic leukaemia nuclear bodies (PML-NBs) (**c**) are shown (see Supplementary Fig. 9f, g for representative images). **d**, Percentages of SA-β-gal-positive cells (top) and protein expression (bottom) in IMR90 cells at different passages. **e**, **f**, Replicative lifespan (**e**) and protein expression (**f**) of IMR90 cells with and without overexpression of ME1 or ME2. Arrows in **e** indicate the onset of senescence. All error bars represent mean ± s.d. ($n = 3$). $^{**}P < 0.01$.

to induce this phenotype (Fig. 3b, c and Supplementary Figs 9b, c, f, g and 10c). By contrast, malic enzyme depletion did not cause cell death (Supplementary Fig. 11a); it induced the expression of p53 target genes implicated in senescence[2,13] but not apoptosis (Supplementary Fig. 11b). These data indicate that downregulation of malic enzymes induces senescence through p53.

We next examined the role of malic enzymes in replicative senescence of normal human cells, a p53-regulated process[1,3]. IMR90 cells were serially passaged in culture until a substantial number of them (~50%) entered senescence. The expression of ME1 remained at comparable levels at different passages, whereas the expression of ME2, which stayed unchanged initially, noticeably declined at the late stage (Fig. 3d). To test whether the decline in ME2 contributes to senescence in this setting, we evaluated the replicative capacity of IMR90 cells forced to express ME2. Compared with control cells, ME2-overexpressing cells could be cultured for extended passages with a greatly delayed onset of senescence (Fig. 3e and Supplementary Fig. 12a). As ME1 expression was maintained during replicative senescence, we were surprised to observe a delay in senescence when ME1 expression was forced (Fig. 3e and Supplementary Fig. 12a). By contrast, forced expression of any of the malic enzyme mutants did not delay senescence and instead moderately promoted senescence (Supplementary Fig. 12b–d), possibly through a dominant-negative effect on the endogenous malic enzymes. These results indicate that both enzymes—particularly ME2—are capable of suppressing senescence and suggest that the decline in ME2 may contribute to replicative senescence.

To examine the effect of malic enzymes on other scenarios of p53-regulated senescence, we found that culturing IMR90 cells in medium containing no or low levels of glutamine resulted in p53-dependent senescence (Supplementary Fig. 13a, b). This senescence could be delayed by overexpression of either ME1 or ME2 (Supplementary Fig. 13b), or the addition of the malic enzyme substrate malate (Supplementary Fig. 13c). By contrast, exogenous malic enzyme expression did not influence premature senescence of IMR90 cells induced by the oncogene *HrasV12* (Supplementary Fig. 13d), which is not dependent on p53 (refs 14, 15). These results indicate that ME1 and ME2 expression suppress the specific way in which p53 induces senescence.

We investigated the mechanism for senescence induced by malic enzyme downregulation. In IMR90 and U2OS cells in which the expression of either *ME1* or *ME2* was silenced by siRNA, even moderately, p53 levels were increased, accompanied by enhanced phosphorylation of p53 and induction of its target gene *p21* (also known as *CDKN1A*) (Figs 2a and 4a and Supplementary Figs 5a, b and 10a). By contrast, overexpression of ME1 or ME2 in IMR90 cells substantially reduced p53 levels and activity in late passages (Fig. 3f). Overexpression of ME1 or ME2 in U2OS cells also diminished DNA-damage-induced p53 activation (Supplementary Fig. 14). These observations suggest a strong role for malic enzymes in the suppression of p53. They also indicate the existence of a positive-feedback loop for the p53–malic enzyme pathway: a higher p53 level leads to less malic enzyme expression, which alleviates the inhibition of malic enzymes on p53, leading to even higher p53 activation.

We next examined the mechanism for the regulation of p53 by ME1 and ME2. In unstressed cells, MDM2-mediated ubiquitination maintains a low basal level of p53 (ref. 2). When *ME1* was knocked down in both IMR90 and U2OS cells, the abundance of the MDM2 protein and mRNA declined markedly (Fig. 4a and Supplementary Fig. 15a, b), suggesting that ME1 downregulation activates p53 through a reduction in MDM2 expression. *ME2* knockdown did not significantly affect MDM2 levels. Instead, it turned on AMP-activated protein kinase (AMPK) (Fig. 4a and Supplementary Fig. 15a, b), an intracellular energy gauge that activates p53 through phosphorylation[16]. We tested whether AMPK is required for induction of p53 by ME2 by knocking it down in IMR90 and U2OS cells, and by comparing *AMPK* null and wild-type MEFs. In both situations, loss of AMPK expression prevented *ME2* knockdown from activating p53 (Fig. 4b and Supplementary Fig. 15c). Because silencing of *ME2* did not influence

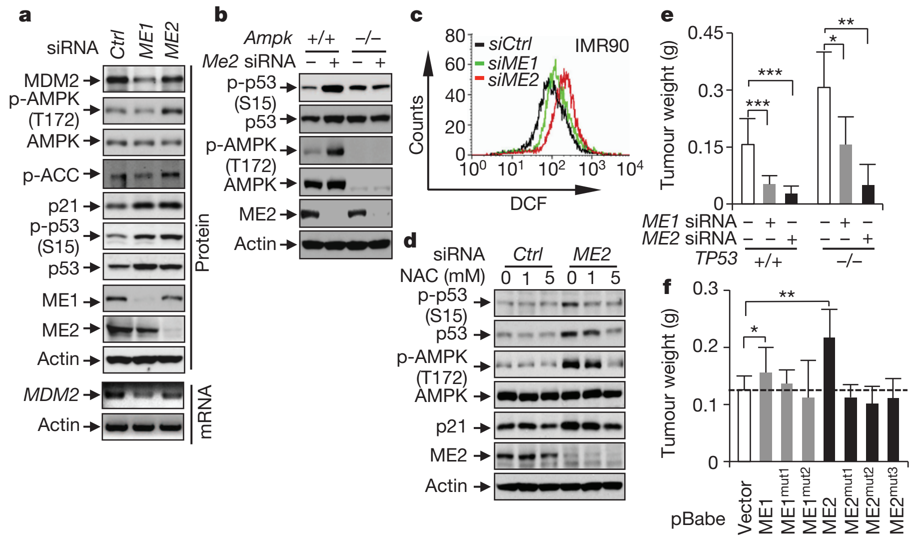


Figure 4 | Mechanisms of p53 activation induced by malic enzyme downregulation and a role of malic enzymes in tumour growth. **a**, Effect of *ME1* and *ME2* knockdown on p53 and AMPK activation and MDM2 expression. ACC, acetyl-CoA carboxylase; p-, phosphorylated. **b**, p53 and AMPK activation in $Ampk^{+/+}$ and $Ampk^{-/-}$ MEF cells transfected with control or *Me2* siRNA. **c**, ROS levels, determined by 2′,7′-dichlorodihydrofluorescein diacetate (DCF), in IMR90 cells transfected with *ME1*, *ME2* or control siRNA. **d**, Effect of *N*-acetyl-L-cysteine (NAC) on AMPK and p53 activation in IMR90 cells transfected with control or *ME2* siRNA. **e**, **f**, Average weights of xenograft tumours (mean ± s.d., $n = 6$) generated by $TP53^{+/+}$ and $TP53^{-/-}$ HCT116 cells transfected with *ME1*, *ME2* or control siRNA (**e**), or $TP53^{+/+}$ HCT116 cells stably overexpressing wild-type or mutant malic enzymes (**f**). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

cellular ATP levels (Supplementary Fig. 8c), we examined other pathways that could activate AMPK and observed a strong increase in reactive oxygen species (ROS) in various *ME2*-depleted cells (Fig. 4c and Supplementary Fig. 16), correlating with a strong reduction in NADPH levels (Fig. 2a and Supplementary Fig. 5a–c). ROS are a known activator of AMPK[17]. Treatment with the ROS scavenger *N*-acetyl-L-cysteine blocked AMPK and p53 activation (Fig. 4d and Supplementary Fig. 17a, b), abrogated senescence and restored growth of *ME2*-knockdown cells (Supplementary Fig. 17c, d). These results indicate that downregulation of *ME2* increases ROS levels leading to sequential activation of AMPK and p53 and the induction of senescence. In comparison, depletion of *ME1* increased ROS levels moderately in IMR90 cells and minimally in U2OS cells (Fig. 4c and Supplementary Fig. 16), and *N*-acetyl-L-cysteine only slightly affected p53 activation, senescence and growth arrest in *ME1*-depeleted cells (Supplementary Fig. 17b–d).

Previous studies on limited tumour samples suggest that the activity of ME2 is highly increased in these tumours and correlates with tumour progression[18–20]. A survey of public gene-expression databases (http://www.oncomine.org) showed that both ME1 and ME2 expression was significantly upregulated in a variety of human cancers (Supplementary Fig. 18). We investigated whether malic enzymes could influence tumour cell growth. Depletion of *ME1* or *ME2* in U2OS and HCT116 cells, regardless of p53 status, strongly impaired their growth (Supplementary Fig. 19), and reduced the number of cells at the S phase of the cell cycle (Supplementary Fig. 20). By contrast, overexpression of ME1 or ME2, but none of the malic enzyme mutants, enhanced tumour cell growth (Supplementary Fig. 21). In a soft agar assay, tumour cells deprived of malic enzyme gene expression, unlike their control counterparts, failed to form anchorage-independent colonies (Supplementary Fig. 22a, b), whereas cells transduced with wild-type malic enzymes, but not any of the mutants, showed enhanced anchorage-independent growth (Supplementary Fig. 22c, d).

To analyse the function of malic enzymes in the tumour xenograft model, we injected immunocompromised mice with $TP53^{+/+}$ and $TP53^{-/-}$ HCT116 cells treated with *ME1*, *ME2* or control siRNA. $TP53^{-/-}$ HCT116 cells gave rise to tumours that were twice the weight of tumours generated by $TP53^{+/+}$ HCT116 cells. When *ME1* or *ME2*—particularly *ME2*—was silenced in these cells, the tumour sizes were markedly reduced (Fig. 4e and Supplementary Fig. 23a, b). $TP53^{+/+}$ HCT116 tumours devoid of malic enzymes showed extensive senescence and were substantially smaller compared to the corresponding $TP53^{-/-}$ HCT116 tumours (Supplementary Fig. 23c). Conversely, overexpression of wild-type ME1 or ME2, not mutant malic enzymes, accelerated the growth of $TP53^{+/+}$ HCT116 tumours (Fig. 4f and Supplementary Fig. 23d). These observations indicate that malic enzymes are essential for tumour growth through both p53-dependent and -independent mechanisms.

Although p53 is able to induce a range of anti-proliferative responses, emerging evidence indicates that senescence induction and metabolic regulation are central to its function as a tumour suppressor[13,21–25]. Our results demonstrate a positive-feedback loop comprising p53 and malic enzymes that influences p53 activation and links metabolism with the onset of senescence (Supplementary Fig. 24). p53 suppresses all malic enzyme (1, 2 and 3) expression by directly binding to response elements within these genes. Together with our recent finding that p53 targets the NADPH producer glucose-6-phosphate dehydrogenase through a distinct direct catalytic mechanism[4], the current study reveals p53 as a master immediate regulator of cellular NADPH levels. p53 is reciprocally regulated by ME1 and ME2. The marked stabilization of p53 upon *ME1* and *ME2* downregulation is achieved through different mechanisms, through the decline of MDM2 levels and ROS-induced AMPK activation, respectively. These findings support the notion that p53 is a central sentinel for metabolic stresses and coordinates metabolic pathways with cell-fate decision.

Mutual regulation of p53 and malic enzymes is likely a key mechanism that modulates cellular senescence in both normal and tumour cells. Even moderate downregulation of either *ME1* or *ME2* strongly induces p53 activation and senescence, whereas overexpression of either enzyme delays these processes. Thus, these enzymes modulate not only the amplitude, but also the outcome, of p53 activation. p53 is subjected to negative-feedback regulation (for example, the p53–MDM2 feedback loop) that restrains its activity[2]. The p53–malic enzyme positive-feedback loop is likely important to alleviate the negative-feedback regulation so that p53 can accumulate to high levels. This may be particularly relevant in situations in which robust and persistent p53 activation is desirable, such as the induction and maintenance

of senescence. The involvement of the p53–malic enzyme pathway in senescence demonstrates a close link between metabolism and this irreversible fate of the cell.

METHODS SUMMARY

Malic-enzyme-dependent glutaminolytic flux was determined by labelling the malate pool with ^{13}C from [U-$^{13}C_5$]glutamine and monitoring the conversion of [^{13}C]malate to pyruvate. ^{13}C enrichments were determined with gas chromatography–mass spectrometry. Glucose and glutamine consumption was determined using a YSI 7100 Multiparameter Bioanalytical System. Detailed experimental procedures are presented in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 5 February; accepted 9 November 2012.
Published online 13 January 2013.

1. Campisi, J. & d'Adda di Fagagna, F. Cellular senescence: when bad things happen to good cells. *Nature Rev. Mol. Cell Biol.* **8,** 729–740 (2007).
2. Vousden, K. H. & Prives, C. Blinded by the light: the growing complexity of p53. *Cell* **137,** 413–431 (2009).
3. Ben-Porath, I. & Weinberg, R. A. The signals and pathways activating cellular senescence. *Int. J. Biochem. Cell Biol.* **37,** 961–976 (2005).
4. Jiang, P. *et al.* p53 regulates biosynthesis through direct inactivation of glucose-6-phosphate dehydrogenase. *Nature Cell Biol.* **13,** 310–316 (2011).
5. Hsu, R. Y. Pigeon liver malic enzyme. *Mol. Cell. Biochem.* **43,** 3–26 (1982).
6. Chang, G. G. & Tong, L. Structure and function of malic enzymes, a new class of oxidative decarboxylases. *Biochemistry* **42,** 12721–12733 (2003).
7. Pongratz, R. L., Kibbey, R. G., Shulman, G. I. & Cline, G. W. Cytosolic and mitochondrial malic enzyme isoforms differentially control insulin secretion. *J. Biol. Chem.* **282,** 200–207 (2007).
8. Riley, T., Sontag, E., Chen, P. & Levine, A. Transcriptional control of human p53-regulated genes. *Nature Rev. Mol. Cell Biol.* **9,** 402–412 (2008).
9. Murphy, M. *et al.* Transcriptional repression by wild-type p53 utilizes histone deacetylases, mediated by interaction with mSin3a. *Genes Dev.* **13,** 2490–2501 (1999).
10. Wellen, K. E. *et al.* ATP-citrate lyase links cellular metabolism to histone acetylation. *Science* **324,** 1076–1080 (2009).
11. Ferbeyre, G. *et al.* PML is induced by oncogenic *ras* and promotes premature senescence. *Genes Dev.* **14,** 2015–2027 (2000).
12. Pearson, M. *et al.* PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. *Nature* **406,** 207–210 (2000).
13. Brady, C. A. *et al.* Distinct p53 transcriptional programs dictate acute DNA-damage responses and tumor suppression. *Cell* **145,** 571–583 (2011).
14. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88,** 593–602 (1997).
15. Wei, W., Hemmer, R. M. & Sedivy, J. M. Role of p14^{ARF} in replicative and induced senescence of human fibroblasts. *Mol. Cell. Biol.* **21,** 6748–6757 (2001).
16. Jones, R. G. *et al.* AMP-activated protein kinase induces a p53-dependent metabolic checkpoint. *Mol. Cell* **18,** 283–293 (2005).
17. Blättler, S. M., Rencurel, F., Kaufmann, M. R. & Meyer, U. A. In the regulation of cytochrome P450 genes, phenobarbital targets LKB1 for necessary activation of AMP-activated protein kinase. *Proc. Natl Acad. Sci. USA* **104,** 1045–1050 (2007).
18. Wasilenko, W. J. & Marchok, A. C. Malic enzyme and malate dehydrogenase activities in rat tracheal epithelial cells during the progression of neoplasia. *Cancer Lett.* **28,** 35–42 (1985).
19. Sauer, L. A., Dauchy, R. T., Nagel, W. O. & Morris, H. P. Mitochondrial malic enzymes. Mitochondrial NAD(P)$^+$-dependent malic enzyme activity and malate-dependent pyruvate formation are progression-linked in Morris hepatomas. *J. Biol. Chem.* **255,** 3844–3848 (1980).
20. Nagel, W. O., Dauchy, R. T. & Sauer, L. A. Mitochondrial malic enzymes. An association between NAD(P)$^+$-dependent malic enzyme and cell renewal in Sprague-Dawley rat tissues. *J. Biol. Chem.* **255,** 3849–3854 (1980).
21. Braig, M. *et al.* Oncogene-induced senescence as an initial barrier in lymphoma development. *Nature* **436,** 660–665 (2005).
22. Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. *Nature* **436,** 725–730 (2005).
23. Xue, W. *et al.* Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature* **445,** 656–660 (2007).
24. Ventura, A. *et al.* Restoration of p53 function leads to tumour regression *in vivo.* *Nature* **445,** 661–665 (2007).
25. Li, T. *et al.* Tumor suppression in the absence of p53-mediated cell-cycle arrest, apoptosis, and senescence. *Cell* **149,** 1269–1283 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. J. Birnbaum, B. Vogelstein, W. El-Deiry and M. Lazar for reagents; M. J. Bennett, S. Patel, A. Stonestrom and M. Brewer for technical assistance; and A. Stonestrom for help with manuscript preparation. This work was supported by grants from the National Institutes of Health (CA088868) and the US Department of Defense (W81XWH-10-1-0468) to X.Y.

Author Contributions P.J., W.D. and X.Y. designed the study, interpreted the data and wrote the manuscript. P.J. and W.D. performed the experiments. K.E.W. helped with the metabolic studies and data interpretation. A.M. designed the glutaminolytic flux procedure and performed the experiment with the help from P.J.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.Y. (xyang@mail.med.upenn.edu).

METHODS

Antibodies and reagents. The antibodies against the following proteins/epitopes were purchased from the indicated sources: ME2, actin and β-tubulin (Sigma); AMPK, phospho-AMPK (Thr 172), phospho-p53 (Ser15) and phospho-acetyl-coenzyme A carboxylase (Ser 79) (Cell Signaling Technology); ME1, p21 and PML (Santa Cruz Biotechnology); prohibitin (Thermo Scientific); GAPDH (Novus Biologicals); p53 (DO-1; Oncogene, and Santa Cruz Biotechnology); and MDM2 (Calbiochem, and Santa Cruz Biotechnology). The following reagents were purchased from Sigma: dimethyl L-malate, $NADP^+$, NAD^+, doxorubicin, etoposide, trichostatin A, NAC, insulin, troglitazone, dexamethasone and isobutylmethylxanthine.

Cell culture and gene knockdown with shRNA and siRNA. Cells were maintained in standard culture conditions without any antibiotic. Expression plasmids for *TP53*, *ME1* and *ME2* shRNA were made in a pLKO.1-puro vector. The targeted sequences for human *TP53* and mouse *Trp53* are 5′-GACTCCAGTGGT AATCTAC-3′ (ref. 26) and 5′-GTACTCTCCTCCCCTCAAT-3′ (ref. 27), respectively. The targeted sequences for human *ME1* and *ME2* are 5′-GGGCAT ATTGCTTCAGTTC-3′ and 5′-GCACGGCTGAAGAAGCATATA-3′, respectively. Stable shRNA cell lines were established as previously described[28]. siRNAs for *ME1*, *ME2* and *AMPK* were purchased form Invitrogen. siRNA sequences were 5′-AUAACAAUCAGGUAGAAUCUGGUCA-3′ (human *ME1*), 5′-UAUAGU UGAAGGCUUCAGUAUAUUC-3′ (human *ME2*), 5′-CCCUGUGGGUAAAU UGGCUCUAUAU-3′ (human *ME1* no. 2), 5′-CCUGACAAGCCAAUUGACAG AUGAA-3′ (human *ME2* no. 2), 5′-CGUUGAAAAUUGCAGUAAA-3′ (mouse *Me1*), 5′-GGGCACUGAUAACAUGGCACUAUUA-3′ (mouse *Me2*) and 5′-ACCAUGAUUGAUGAUGAAGCCUUAA-3′ (human *AMPK*). siRNAs were transfected into cells using Lipofectamine RNAiMAX Transfection Agent (Invitrogen).

Semi-quantitative RT–PCR and quantitative RT–PCR. Total RNA was isolated from cells by TRIzol Reagent (Invitrogen). Two micrograms of RNA for each sample were reversed to complementary DNA by First-strand cDNA Synthesis System (Marligen Biosciences), and 0.2 μg cDNA was used as a template to perform PCR. The primer pairs for human genes were: *ME1*, 5′-ACAGATAATAT TTTCCTCACT-3′ and 5′-CTACTGGTCAACTTTGGT-3′; *ME2*, 5′-ATTAGT GACAGTGTTTTCCTA-3′ and 5′-CTATTCTGTTATCACAGG-3′; *p21*, 5′-CCGGCGAGGCCGGGATGAG-3′ and 5′- CTTCCTCTTGGAGAAGATC-3′; *ACTB*, 5′-GACCTGACTGACTACCTCATGAAGAT-3′ and 5′-GTCACACTT CATGATGGAGTTGAAGG-3′; *TP53*, 5′-CACGAGCTGCCCCCAGG-3′ and 5′-TCAGTCGACGTCTGAGT-3′. Primer pairs for mouse genes were: *Me1*, 5′-GATGATAAGGTCTTCCTCACC-3′ and 5′-TTACTGGTTGACTTTGGTCTGT-3′; *Me2*, 5′-TTCTTAGAAG CTGCAAAGGC-3′ and 5′-TCAGTGGGGAAGCT TCTCTT-3′; *p21*, 5′-AACTTCGTCTGGGAGCGC-3′ and 5′-TCAGGGTTTTCT CTTGCAGA-3′; *Actb*, 5′-ACTACATTCAATTCCATC-3′ and 5′-CTAGAAGC ACTTGCGGTG-3′; *Trp53*, 5′-GAAGTCCTTTGCCCTGAAC-3′ and 5′-CTAGC AGTTTGGGCTTTCC-3′.

All RT–PCR reactions were performed using the 7900HT Fast Real-Time PCR System (Applied Biosystems) and the amplifications were done using the SYBR Green PCR Master Mix (Applied Biosystems). The thermal cycling conditions were: 50 °C for 2 min followed by an initial de-naturation step at 95 °C for 10 min, 45 cycles at 95 °C for 15 s, 60 °C for 1 min, and a dissociation curve at 95 °C for 15 s and 60 °C for 15 s. The experiments were carried out in triplicate for each data point. Using this method, we obtained the fold changes in gene expression normalized to an internal control gene.

Cell lysate fractionation and malic enzyme activity. Cell fractionation was carried out as described[29]. Cells were homogenized in 20 mM HEPES-KOH buffer, pH 7.5, 10 mM KCL, 1.5 mM $MgCl_2$, 1 mM sodium EDTA buffer, 1 mM sodium EGTA buffer and 1 mM dithiothreitol in the presence of 250 mM sucrose and protease inhibitor cocktail (Roche Diagnostics). Homogenates were centrifuged at 500*g* for 5 min at 4 °C, and the supernatant was collected and centrifuged again at 10,000*g* for 20 min to obtain cytosolic and mitochondrial fractions.

ME1 activity was determined using cytosolic extracts as described[30]. The reaction buffer contained 67 mM triethanolamine, 3.3 mM L-malic acid, 0.3 mM β-$NADP^+$ and 5.0 mM manganese chloride. For measuring ME2 activity, mitochondria were purified as described[29] and re-suspended in mitochondrial lysis buffer (20 mM MOPS-KOH, pH 7.4, 250 mM sucrose, 80 mM KCl, 5 mM EDTA, 1 mM PMSF, 1% Triton X-100 and protease inhibitor cocktail) on ice for 30 min by gentle vortexing for 5 s at 5-min intervals. Lysates were centrifuged for 10 min at 14,000 r.p.m. at 4 °C. The enzyme reaction mixtures contained 50 mM Tris-HCl, pH 7.4, 10 mM $MgCl_2$, 0.3 mM NAD^+ and 3.3 mM L-malic acid. The reactions were started by adding either cytosolic and mitochondrial extracts, and were monitored by absorbance at 340 nm every 5 s for up to 10 min. Background control was run without L-malic acid as substrate. Enzyme activity was determined by subtracting the activity of the background control to each sample. The resulting changes of absorbance versus time were normalized to protein contents, which were determined using the Bio-Rad protein assay.

Analysis of malic enzyme gene sequence and chromatin immunoprecipitation (ChIP) assay. We used the Genomatix Promoter Inspector software (http://www.genomatix.de) to search in malic enzyme genes for potential p53 response elements with the consensus sequence 5′-RRRCWWGYYY-(0–13-base pair spacer)-RRRCWWGYYY-3′, in which R is a purine, Y a pyrimidine, and W either A or T[8]. The sequences for the putative p53 response elements in malic enzyme genes are: ME1-RE, 5′-TTACCTGGTTAACTAGGACTTGCCC-3′; ME2-RE1, 5′-AGGCATGCACCACCATGCCC-3′; ME2-RE2, 5′-AGACCAGTCAAAAAC ATGTCC-3′; ME2-RE3, 5′-GGGCATGATGGCACATGCCT-3′; and ME3-RE, 5′-TGACTTGGTTTGGCTTTCTTGTCC-3′.

For ChIP assays, cells were washed with PBS and crosslinked with a 1% formaldehyde solution for 15 min at room temperature (25 °C). The crosslinking reaction was stopped by the addition of glycine to 125 mM final concentration. Cell lysates were sonicated to generate DNA fragments with the average size below 1,000 base pairs and followed by immunoprecipitation with indicated antibodies. Bounded DNA fragments were eluted and amplified by PCR. The primer pairs were: ME1-RE, 5′-GCCTTAGTATGTGGATTC-3′ and 5′-GGAAAGCGTAG GGAAGGA-3′; ME2-ME1, 5′-GTTGCCCAGGCTGGAGTG-3′ and 5′-CTGT AATCCCAGCACTTT-3′; ME2-RE2, 5′-TCAGCACTTTGGGAGG-3′ and 5′-GCGACAGAGTCTTGCC-3′; ME2-RE3, 5′-GGCTCAGTGGCTCACG-3′ and 5′-GTGCAGTGGCATG-3′; ME3-RE, 5′-GTTGCGATCCCGTGGCTG-3′ and 5′-ACCGCAGGTCAGACTGAC-3′; p21, 5′-CTGAAAACAGGCAGCCCAAG-3′ and 5′-GTGGCTCTGATTGGCTTTCTG-3′[28].

Reporter assay. The DNA fragment containing the potential p53-binding region was amplified by PCR with primers used in the ChIP assay and was cloned into a pGL3-promoter vector (Promega). 293T cells were plated 18 h before transfection in 24-well plates and transiently transfected with 450 ng of the reporter plasmid and/or 100 ng of the p53 plasmid using Lipofectamine 2000 (Invitrogen). The luciferase activity was determined according to the manufacturer's instructions (Promega). Transfection efficiency was normalized on the basis of the Renilla luciferase activity.

Measurements of metabolites and lipid accumulation. The levels of NADPH, NADH, NAD^+, ATP and ADP in cultured cells were determined using a $NADP^+$/NADPH Quantification Kit, NAD^+/NADH Quantification Kit, ATP assay kit, and ADP assay kit (all from BioVision) respectively, following the manufacturer's instructions. Glucose and glutamine consumption was determined using YSI 7100 Multiparameter Bioanalytical System (YSI Life Sciences). Triglyceride was measured using a Triglyceride Assay Kit (BioVision). Total lipids were measured using Oil Red O staining[10]. For this, confluent cells were grown in medium with 10% FBS supplemented with insulin ($5\ \mu g\ ml^{-1}$), dexamethasone (1 μM), troglitazone (5 μM) and isobutylmethylxanthine (0.5 mM) for 2 days, and in medium supplemented with insulin and rosiglitazone for an additional 5 days. The medium was changed every other day. Cells were then fixed with 4% paraformaldehyde for 30 min at room temperature, washed with distilled H_2O and 60% isopropanol, and stained with a filtered Oil Red O work solution at room temperature. Stain was then removed and cells were washed four times in distilled H_2O.

Measurement of glutaminolytic flux. Malic-enzyme-dependent glutaminolytic flux was determined by labelling the malate pool with ^{13}C from [U-$^{13}C_5$]glutamine and monitoring the conversion of malate to pyruvate indirectly by detecting ^{13}C-labelled lactate. Carbon-13 enrichments were determined with gas chromatography–mass spectrometry. We observed that this approach results in approximately 50% ^{13}C enrichment and allows for the determination of glutaminolytic flux through malic enzyme with high sensitivity. Cells cultured on 10-cm plates with around 60% confluence were transfected with control, *ME1* or *ME2* siRNA. Cells were cultured in regular medium for 60 h and in DMEM containing 12 mM glucose, 3 mM [U-$^{13}C_5$]glutamine and 10% dialysed FBS (Sigma) for an additional 9 h. After medium was removed, cells were immediately quenched with cold 80% methanol. The methanol/cell mass mixtures were centrifuged. The insoluble (protein and lipid) fraction was analysed for protein content. The soluble fraction was dried under a stream of gaseous nitrogen at 40 °C and silylated with *N-tert*-Butyldimethylsilyl-*N*-methyltrifluoroacetamide (MTBSTFA; Regis). The silylated cell extracts were analysed with an Agilent 7890A gas chromatograph/5975C mass spectrometer (Agilent). Mass spectra were quantified with the MSD ChemStation software (Agilent) and corrected for natural abundance contributions from ^{13}C, ^{29}Si and ^{30}Si using Isocor (http://www.python.org)[31]. The total lactate level was determined with the YSI 7100 Multiparameter Bioanalytical System. The glutaminolytic flux through malic enzymes was calculated from the equation: $F_{ME} = F_L \times (L_{m+3})/(M_{m+4})$, in which F_{ME} = malic enzyme flux, F_L = total lactate flux, L_{m+3} = fraction of lactate enriched in all three carbons, and M_{m+4} = fraction of malate enriched in all four carbons.

Measurements of ROS. ROS levels were determined as described[32]. Cells were incubated at 37 °C for 30 min in PBS containing 10 μM 2',7'-dichlorodihydrofluorescein diacetate (H2-DCFDA, Sigma). Afterwards, the cells were washed twice in PBS, treated with trypsin, and re-suspended in PBS. Fluorescence was immediately measured using a FACScan Flow Cytometer (Becton Dickinson).

Senescence-associated SA-β-gal activity. The SA-β-gal activity in cultured cells was determined using a Senescence Detection Kit (BioVision) following the manufacturer's instructions. Percentages of cells that stained positive were calculated by counting 1,000 cells in random fields per cell line.

Immunofluorescence. Cells treated with siRNA for 48 h were washed with 1xPBS and fixed in 4% paraformaldehyde. After being treated with 0.1% Triton X-100, cells were stained using anti-PML antibody followed by Texas-red conjugated anti-mouse IgG antibody, and mounted with 4,6-diamidino-2-phenylindole (DAPI) (Vector Laboratories). The images were acquired with a confocal microscope. A total of 200 nuclei were selected randomly and promyelocytic leukaemia nuclear bodies within each nucleus were counted.

Cell proliferation assay. Cells were treated with siRNAs for 24 h and seeded in 6-well cell culture dishes in triplicates at a density of 20,000 cells per well in 2 ml of medium containing 10% FBS. The medium was changed everyday. Cells were counted and cell number at the indicated time points was determined.

Soft agar assay and xenograft tumour models. For the soft agar assay, cells were suspended in 1 ml of 10% FBS DMEM medium containing a 0.3% agarose and plated on a firm 0.6% agarose base in 6-well plates (5,000 cells per well) as described previously[33]. Cells were then cultured in a 37 °C and 5% CO_2 incubator for 2 weeks. Images were obtained and colonies were counted under a microscope. Each experiment was done in triplicate. For the mouse xenograft experiment, cells (2×10^6) were injected subcutaneously into the flanks of 4- to 5-week-old athymic Balb-c nu/nu male mice (Taconic Farms). Tumour growth was evaluated at 2 weeks post-injection. All animal experiments were performed in accordance with relevant guidelines and regulations and were approved by the University of Pennsylvania Institutional Animal Care and Use Committee (IACUC).

26. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296,** 550–553 (2002).
27. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proc. Natl Acad. Sci. USA* **101,** 10380–10385 (2004).
28. Godar, S. *et al.* Growth-inhibitory and tumor-suppressive functions of p53 depend on its repression of *CD44* expression. *Cell* **134,** 62–73 (2008).
29. Jiang, P., Du, W., Heese, K. & Wu, M. The Bad guy cooperates with good cop p53: Bad is transcriptionally up-regulated by p53 and forms a Bad/p53 complex at the mitochondria to induce apoptosis. *Mol. Cell. Biol.* **26,** 9071–9082 (2006).
30. Guay, C., Madiraju, S. R., Aumais, A., Joly, E. & Prentki, M. A role for ATP-citrate lyase, malic enzyme, and pyruvate/citrate cycling in glucose-induced insulin secretion. *J. Biol. Chem.* **282,** 35657–35665 (2007).
31. Millard, P., Letisse, F., Sokol, S. & Portais, J. C. IsoCor: correcting MS data in isotope labeling experiments. *Bioinformatics* **28,** 1294–1296 (2012).
32. Cossarizza, A. *et al.* Simultaneous analysis of reactive oxygen species and reduced glutathione content in living cells by polychromatic flow cytometry. *Nature Protocols* **4,** 1790–1797 (2009).
33. Zhang, J. *et al.* AFAP-110 is overexpressed in prostate cancer and contributes to tumorigenic growth by regulating focal contacts. *J. Clin. Invest.* **117,** 2962–2973 (2007).

doi:10.1038/nature11779

Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence

Yuval Tabach[1,2], Allison C. Billi[3,4]*, Gabriel D. Hayes[1,2]*, Martin A. Newman[1,2], Or Zuk[5], Harrison Gabel[1,2], Ravi Kamath[1,2], Keren Yacoby[1], Brad Chapman[1], Susana M. Garcia[1,2], Mark Borowsky[1,2], John K. Kim[3,4] & Gary Ruvkun[1,2]

Genetic and biochemical analyses of RNA interference (RNAi) and microRNA (miRNA) pathways have revealed proteins such as Argonaute and Dicer as essential cofactors that process and present small RNAs to their targets. Well-validated small RNA pathway cofactors such as these show distinctive patterns of conservation or divergence in particular animal, plant, fungal and protist species. We compared 86 divergent eukaryotic genome sequences to discern sets of proteins that show similar phylogenetic profiles with known small RNA cofactors. A large set of additional candidate small RNA cofactors have emerged from functional genomic screens for defects in miRNA- or short interfering RNA (siRNA)-mediated repression in *Caenorhabditis elegans* and *Drosophila melanogaster*[1,2], and from proteomic analyses of proteins co-purifying with validated small RNA pathway proteins[3,4]. The phylogenetic profiles of many of these candidate small RNA pathway proteins are similar to those of known small RNA cofactor proteins. We used a Bayesian approach to integrate the phylogenetic profile analysis with predictions from diverse transcriptional coregulation and proteome interaction data sets to assign a probability for each protein for a role in a small RNA pathway. Testing high-confidence candidates from this analysis for defects in RNAi silencing, we found that about one-half of the predicted small RNA cofactors are required for RNAi silencing. Many of the newly identified small RNA pathway proteins are orthologues of proteins implicated in RNA splicing. In support of a deep connection between the mechanism of RNA splicing and small-RNA-mediated gene silencing, the presence of the Argonaute proteins and other small RNA components in the many species analysed strongly correlates with the number of introns in those species.

Proteins with similar patterns of conservation or divergence across different organisms are more likely to act in the same pathways[5]. To identify proteins that share an evolutionary history with validated small RNA pathway proteins, we determined the phylogenetic profiles of approximately 20,000 *C. elegans* proteins in 85 genomes, representing diverse taxa of the eukaryotic tree of life: 33 animals, 6 land plants, 1 alga, 31 Ascomycota fungi, 3 Basidiomycota fungi and 12 protists. Of the ~20,000 *C. elegans* proteins, 10,054 show homologues in non-nematode eukaryotic genomes (Supplementary Table 1). Following correlation and clustering, this analysis sorts genes into clades of conservation and relative divergence or loss in the various organisms as suites of genes are maintained from common ancestors or diverge in particular lineages[6]. Protein divergence or loss in particular taxonomic clades is not random; entire suites of proteins can diverge or be lost as particular taxa specialize and no longer require ancestral functions. The correlated loss of proteins has been used to assign roles for nuclear-encoded mitochondrial proteins[7] and eukaryotic cilia-associated proteins[8].

We developed a non-binary method of phylogenetic profiling to cluster all protein sequences encoded by *C. elegans* genes. BLAST scores were normalized to the length of the query sequence and for relative phylogenetic distance between *C. elegans* and the queried organism[9]. The matrix of 864,644 conservation scores for the 10,054 *C. elegans* proteins in the 86 genomes was queried either with a single protein to generate a ranking of other *C. elegans* proteins with the most similar pattern of conservation values or using a more global hierarchical clustering method (Fig. 1a). Proteins of the same families exhibit similar patterns of phylogenetic conservation and therefore tend to group together in the hierarchical clustering. However, many phylogenetic clusters include proteins with no sequence similarity; only their conservation or divergence in genomes is correlated. The ability of this non-binary method of phylogenetic profiling to cluster proteins based on function is exemplified by the clustering of proteins known to act as members of complexes. For example, the known protein components of the sensory cilium have highly correlated phylogenetic profiles characterized by loss in particular vertebrates and all fungi and plants and retention in particular protists, whereas the extraordinarily high and universal conservation of ribosomal and translation factor proteins clusters many of these translation components (Supplementary Fig. 1a, b).

With a simple query of one of the central proteins in RNAi, the Argonaute RDE-1, we generated a rank-ordered list of proteins with phylogenetic profiles most similar to that of RDE-1 (Fig. 1b). The 26 other *C. elegans* Argonautes represent the top correlated proteins, a trivial consequence of protein sequence similarity within the Argonaute family. The signature phylogenetic profile of the Argonaute proteins is that they are absent in 9 out of 31 Ascomycota species, 1 out of 3 Basidiomycota species, and 6 out of 14 protist species, but have not been lost in any of the 33 animal or 6 land plant species compared. The retention of Argonaute proteins correlates with the ability to inactivate genes by RNAi[10], and the loss of RNAi in about one-half of the sequenced Ascomycota fungi is correlated with the 'killer' RNA virus[11]. Additional *C. elegans* proteins that cluster with the Argonautes but show no sequence similarity include an asparaginase encoded by *K01G5.9*, the CAND-1 elongation factor and another elongation factor, the THO complex protein THOC-1. THO complex members have emerged from genetic screens for defective transgene and RNAi silencing in *Arabidopsis thaliana*[12].

Another validated *C. elegans* RNAi protein is MUT-2, a polyA polymerase implicated in a step downstream of the production of primary siRNAs by Dicer[13]. Out of the 50 *C. elegans* proteins with phylogenetic profiles most closely correlated with MUT-2 (Supplementary Fig. 1c), 10 are Argonautes, which bear no sequence similarity to MUT-2, demonstrating the efficacy of this approach to detect validated small

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109, USA. [4]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.
*These authors contributed equally to this work.

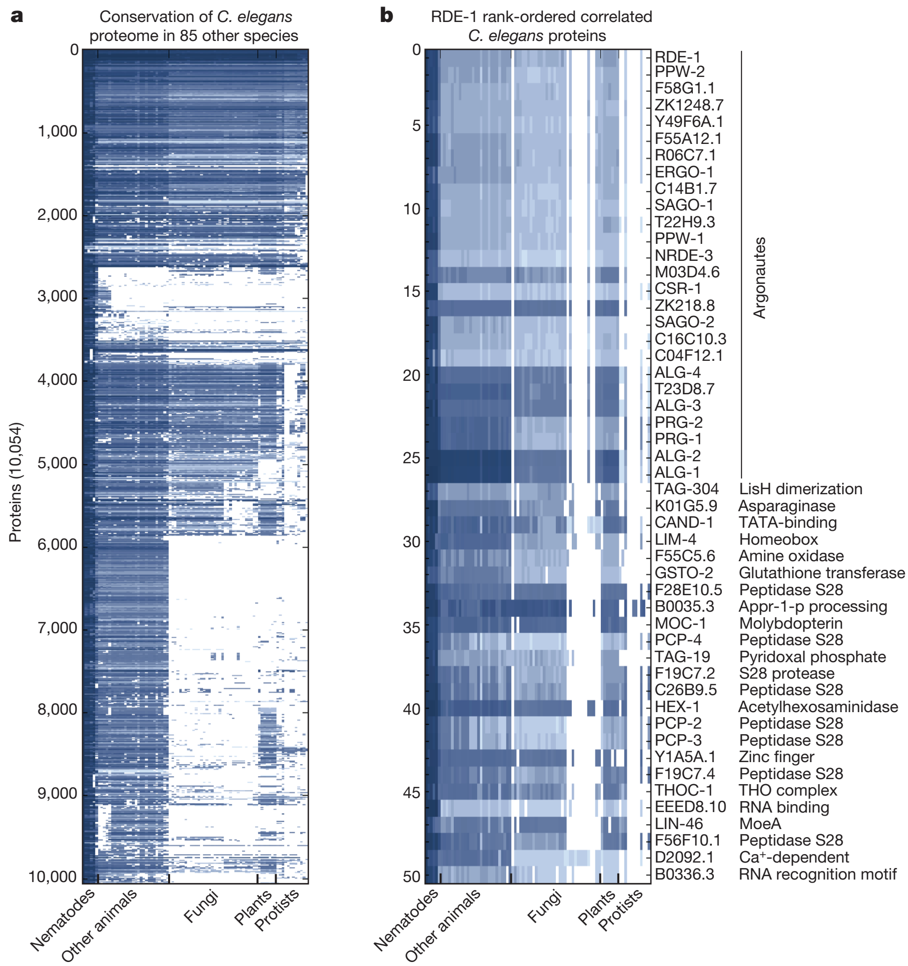


Figure 1 | Phylogenetic profiling analysis shows correlated conservation patterns of *C. elegans* proteins. **a**, Phylogenetic profiles of 10,054 conserved *C. elegans* proteins across 85 other eukaryotic genomes. For each *C. elegans* query protein, the normalized ratio of the BLAST score for the top-scoring protein sequence similarity is indicated in the column corresponding to each genome. Values range from 0 (white, no similarity) to 1 (blue, 100% similarity). **b**, Phylogenetic profiles of validated RNAi factor RDE-1 and the 49 most correlated proteins in rank order.

RNA pathway proteins. The splicing components MAG-1, RSP-8, RNP-4, RSP-5 and DDB-1 and the translation factors EIF-3.D and EIF-3.E, many of which score in the validation tests below, also have similar phylogenetic profiles. In addition, out of the proteins most correlated with the *C. elegans* orthologue of Dicer (DCR-1), a nuclease that processes siRNAs and miRNAs, 3 Argonaute proteins emerge among the top 50 correlated phylogenetic profiles (Supplementary Fig. 1d and Supplementary Table 2).

The RNA-dependent RNA polymerases[14], siRNA-amplifying cofactors, are present in only 5 out of 27 animals (all the nematode species and, surprisingly, the tick), in all of the land plants, in 2 out of 4 Basidiomycota fungi, in 18 out of 27 Ascomycota fungi and in 4 out of 14 protists, but are not present in green algae. A query of the RNA-dependent RNA polymerase RRF-3 (Supplementary Fig. 1e) revealed the cofactor-independent phosphoglycerate mutase F57B10.3 as a dramatically correlated non-homologous protein ($R = 0.93$). Inactivation of this phosphoglycerate mutase gene causes defects in the endogenous siRNA response as well as transgene silencing, validating its role in RNA silencing (Supplementary Table 2). It is possible that either the biochemical substrate or product of this glycolysis pathway protein, or its enzymatic activity as a phosphatase, couples it to small RNA pathways.

To identify candidate small RNA pathway proteins more comprehensively, we globally ranked proteins based on phylogenetic-profile correlation with multiple validated siRNA and miRNA cofactors. After assigning all conserved *C. elegans* proteins to hierarchical clusters, we gave each protein a score to reflect its phylogenetic clustering with the validated set of small RNA proteins (Supplementary Fig. 2). This analysis identified 60 proteins not previously implicated in small RNA pathways whose phylogenetic profiles correlate highly with those of validated siRNA and miRNA pathway proteins (Fig. 2).

The validated siRNA and miRNA protein cofactors identified so far probably constitute a small fraction of the total number of proteins that mediate small RNA function. Full-genome RNAi screens for defects in siRNA or miRNA pathway function have identified hundreds of additional candidate small RNA pathway proteins. We integrated ten genome-scale studies into the phylogenetic cluster analysis: five *C. elegans* gene-inactivation screens for defects in RNAi or miRNA function[1,15,16], *C. elegans* orthologues of *Drosophila* genes identified in two full-genome RNAi screens for impaired siRNA or miRNA response[2] and three proteomic studies of complexes containing the known RNAi proteins DCR-1 (ref. 4), ERI-1 (ref. 17) and AIN-2 (ref. 18). Candidate genes identified in these studies show little overlap (Supplementary Table 3 and Supplementary Fig. 3a, b). However, the candidates from the different studies have similar phylogenetic profiles to each other and to validated small RNA cofactors (Fig. 3, Supplementary Fig. 3c, d and Supplementary Table 4).

We used a naive Bayesian classifier to assign predictive values to six genome-scale studies of RNAi cofactors and five miRNA cofactors (see Supplementary Methods)[19,20]. To the phylogenetic profiles, we added a score for each *C. elegans* gene that is co-expressed on microarrays[21] or whose encoded gene product interacts with validated small RNA pathway proteins[22]. The top 105 genes identified by this analysis are

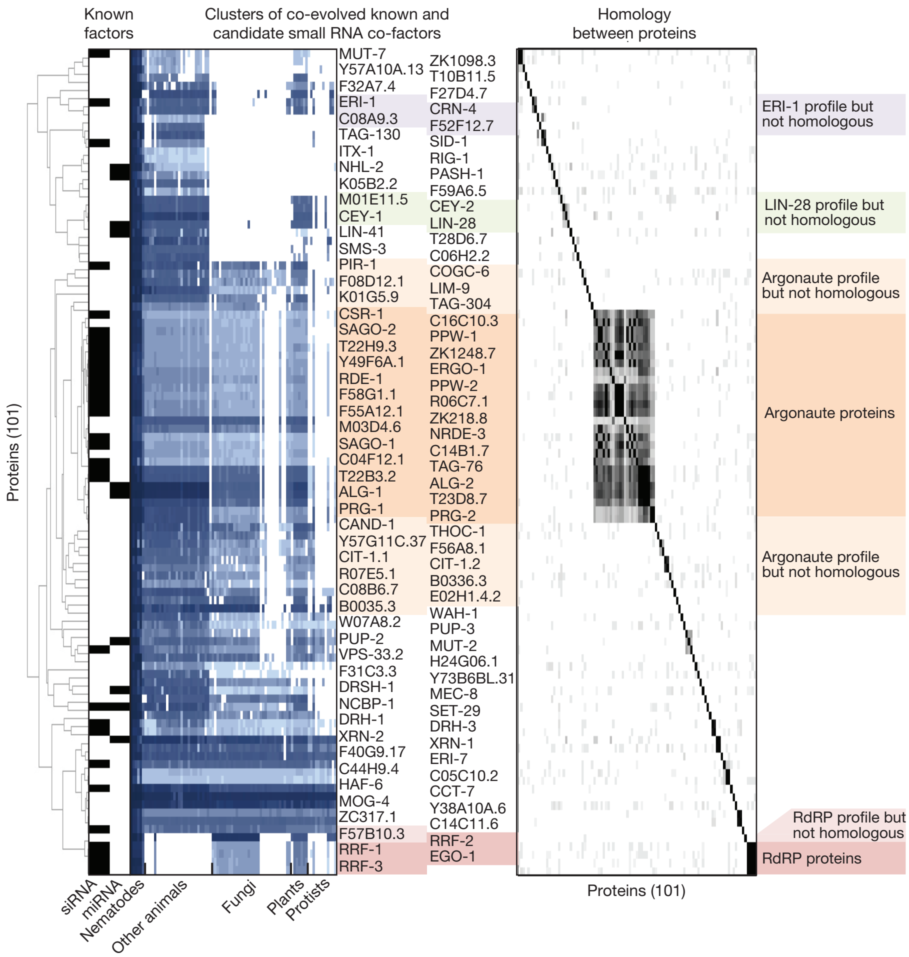


Figure 2 | Phylogenetic clusters of candidate small RNA pathway proteins. Validated miRNA and siRNA pathway proteins map non-randomly on the phylogenetic profile; proteins that map to the same clusters are likely to function in small RNA pathways. Left panel, clusters enriched for validated miRNA and siRNA pathway proteins (black boxes). Darker blue, higher protein-sequence similarity. Right panel, pairwise local protein-sequence alignment of all pairs of proteins in the cluster. White, no similarity; black, significant similarity.

enriched with 41 well-validated siRNA pathway genes (Supplementary Fig. 7 and Supplementary Table 2). The other genes on this list are excellent candidates to mediate siRNA or related small RNA functions. More than 20 of these genes encode RNA recognition motifs including RNP ($P < 0.00001$) and helicase ($P < 0.00001$), an approximately 20-fold enrichment relative to the entire data set. Nine proteins from this list constitute components of the spliceosome (Supplementary Fig. 3).

From the proteins best correlated with validated small RNA pathway cofactors by phylogenetic profile or in the naive Bayesian analysis (Figs 1–3), we tested 87 representative candidates using two different tests for defects in RNAi. Transgene silencing in the somatic cells of the enhanced RNAi mutant *eri-1*(*mg366*) is mediated by an RNAi mechanism[1]. We tested a set of 87 predicted small RNA pathway genes using this strain, and 43 scored as significantly RNAi-defective (Supplementary Table 2, and Fig. 4a). We also tested candidates using a green fluorescent protein (GFP)-based sensor for the abundant *C. elegans* endogenous siRNA 22G siR-1 (ref. 23) to monitor whether any of the gene inactivations affect the production or response to this endogenous siRNA. Thirty-three out of 87 genes tested scored in this assay (Supplementary Table 2 and Fig. 4b). Eight of the nine predicted splicing components scored strongly in these validation screens.

The enrichment for RNA splicing components (Supplementary Fig. 4) points to a close mechanistic connection between splicing and small RNA regulation. Among the Ascomycota and protist species that have lost the Argonaute proteins, most show an extreme loss of introns, from 10^4–10^5 introns in species with Argonautes to 10^2 or fewer introns in most species without Argonautes (Supplementary Fig. 5). We screened for defects in RNAi a cherry-picked gene inactivation sublibrary of *C. elegans* orthologues of known splicing factors that have emerged from biochemical and genetic screens for splicing components from other systems. From a set of 46 *C. elegans* genes annotated in KEGG (Kyoto Encyclopedia of Genes and Genomes) to encode the orthologues of known splicing proteins that could be tested for roles in RNAi in our assays, 16 and 22 of these splicing-factor genes scored strongly in the *eri-1* transgene desilencing assay and the endogenous 22G siR-1 sensor assay. Many of the splicing components that scored strongly in these screens show a phylogenetic profile similar to the Argonaute proteins (Supplementary Fig. 6 and Supplementary Table 6). However, a subset of splicing factors that are well conserved across phylogeny also scored strongly in these assays.

We used the *eri-1* transgene desilencing system to conduct a full-genome screen for gene inactivations that disable transgene silencing and identified 855 genes required for transgene silencing, with more than 200 scoring above 3 on a scale of 0 to 4 for desilencing (Supplementary Table 7). Among gene inactivations that caused the greatest desilencing, 11% correspond to the highest ranked predictions from the siRNA naive Bayesian analysis, a 30-fold enrichment ($P = 4.7 \times 10^{-13}$ using a hypergeometric test) for positives. Out of the 84 splicing factors that have been assigned to specific splicing steps, 49 scored in the full genome screen as required for transgene silencing, and 32 showed phylogenetic profiles clustering with known small RNA factors. The splicing factors that couple to small RNA pathways were not isolated to any particular step of RNA splicing. Splicing factor mutations in *Schizosaccharomyces pombe* disrupt the RNAi-based

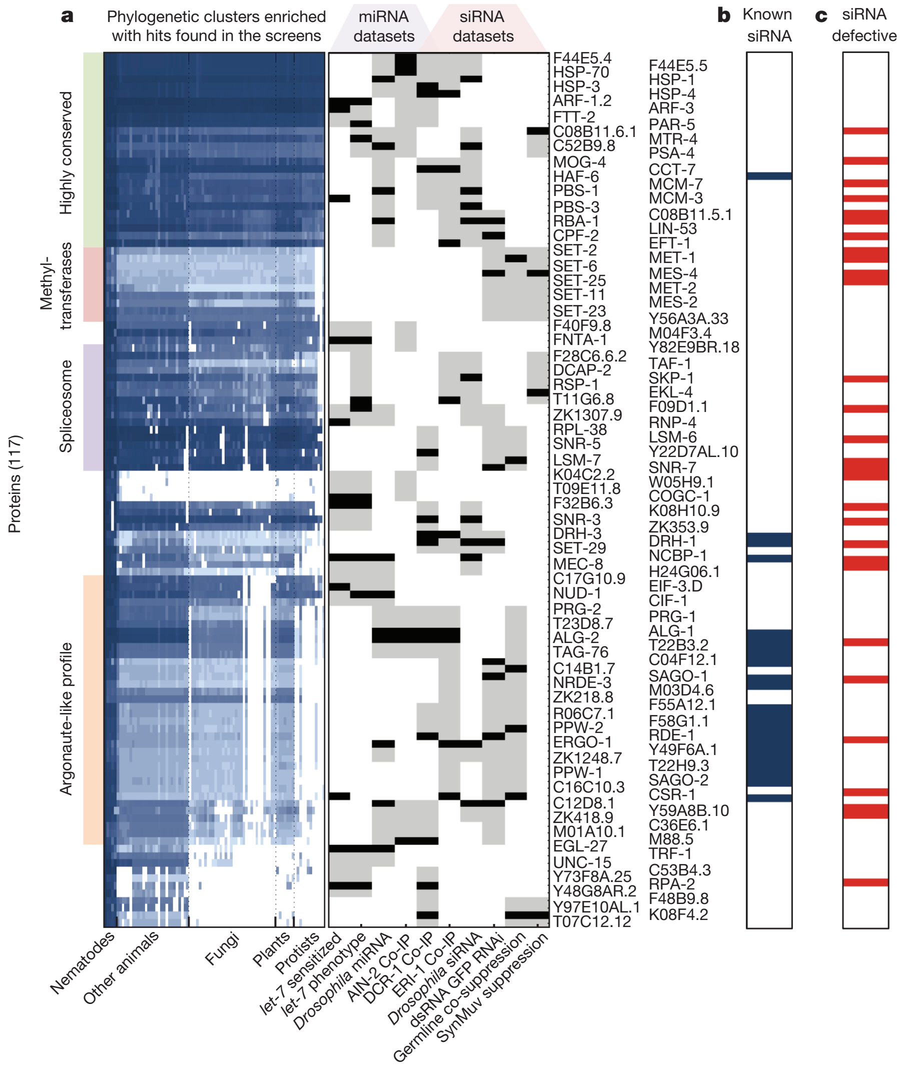


Figure 3 | Select phylogenetic clusters enriched with hits from proteomic and functional genomic small RNA screens. **a**, The phylogenetic profile matrix was clustered and a Max Ratio score (MRS) was calculated for every protein in each screen; 117 proteins scored significantly in miRNA (56 proteins) or siRNA (75 proteins) functional genomic screens, or both (14 proteins). Middle panel, black tick, hit in screens; grey tick, significant MRS. **b**, Blue boxes, the 23 known small RNA pathway proteins identified. **c**, From the 117 proteins predicted by the phylogenetic profile, 28 proteins (red boxes) show defects in siRNA silencing ($P < 3 \times 10^{15}$).

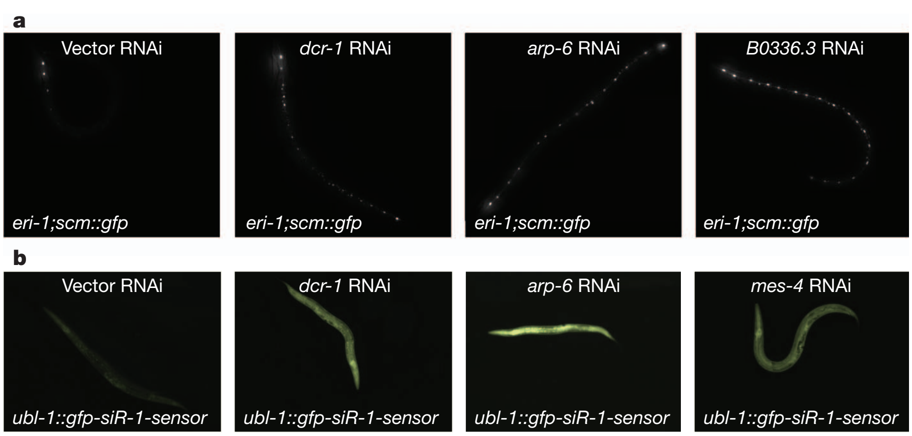


Figure 4 | Inactivation of genes implicated in RNAi pathways re-animates transgenes that are silenced by RNAi. **a**, Expression of *scm::gfp* in the seam cells of an *eri-1(mg366)* mutant, where it is normally silenced by RNAi. Animals shown were treated with control, *dcr-1*, *arp-6* or *B0336.3* RNAi. **b**, GFP expression from the *ubl-1::gfp-siR-1-sensor* transgene, which is normally silenced by the siR-1 endogenous siRNA. Animals shown were treated with control, *dcr-1*, *arp-6* or *mes-4* RNAi.

centromeric silencing[24]. Both splicing proteins and siRNA and miRNA pathway proteins co-localize to cytoplasmic processing bodies (P-bodies) and nuclear Cajal bodies[25], further supporting the possibility of functional crosstalk between splicing and RNAi.

Early genome sequence comparisons of *S. pombe*, *Saccharomyces cerevisiae* and a small set of eukaryotes suggested that loss of introns and splicing components is highly correlated with loss of Argonaute proteins[26]. One interpretation was that the loss of RNAi in *S. cerevisiae* enabled viral invasion and a subsequent loss of introns through reverse transcription of genes by the invading viral replication enzymes. However, such a scenario would not predict that inactivation of splicing components in a species bearing the RNAi apparatus would cause an RNAi-defective phenotype. One model is that splicing could regulate RNAi indirectly by modulating spliced isoforms of key RNAi factors. However, the observations that only a subset of splicing cofactors are required for RNAi and the co-immunoprecipitation of splicing factors and DCR-1, ERI-1 and AIN-2 disfavours this indirect model. A mechanistic coupling between RNAi and RNA splicing explains these new data better. RNAi factors also affect splicing: Dicer is required for efficient spliceosomal RNA maturation in *Candida albicans*[27]. If RNAi engages introns intimately by, for example, engaging nascent transcripts through the Argonaute NRDE-3 before splicing[28], then the selective advantage of introns may fade once the RNAi pathway is lost.

Our data suggest that a large subset of the proteins that mediate steps in the maturation of mRNAs bearing introns are also required for RNAi, and that those genomes that have lost most of their introns no longer require the RNAi pathway. Superimposed on the mRNA splicing pathway is an RNA surveillance system that eliminates aberrantly processed or mutant pre-mRNAs and mRNAs. It is possible that RNAi constitutes another level of mRNA surveillance that acts in parallel to—and using many of the same components as—the splicing quality control surveillance pathways.

METHODS SUMMARY

Informatics. The Normalized Phylogenetic Profile (NPP) data matrix was clustered through MATLAB statistical toolbox using the average linkage method and Pearson correlation coefficient as a similarity measure. Clustering was performed on the rows of the matrix. To identify *C. elegans* proteins with phylogenetic profiles similar to published small RNA co-factors (Supplementary Table 9), the fraction of the validated proteins in each phylogenetic cluster was calculated and optimized to define a Max Ratio Score (MRS) (Supplementary Fig. 2).

Received 16 April; accepted 8 November 2012.
Published online 23 December 2012.

1. Kim, J. K. *et al.* Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308,** 1164–1167 (2005).
2. Zhou, R. *et al.* Comparative analysis of argonaute-dependent small RNA pathways in *Drosophila*. *Mol. Cell* **32,** 592–599 (2008).
3. Meister, G. *et al.* Identification of novel argonaute-associated proteins. *Curr. Biol.* **15,** 2149–2155 (2005).
4. Duchaine, T. F. *et al.* Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. *Cell* **124,** 343–354 (2006).
5. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96,** 4285–4288 (1999).
6. Gabaldón, T. Evolution of proteins and proteomes: a phylogenetics approach. *Evol. Bioinform. Online* **1,** 51–61 (2005).
7. Pagliarini, D. J. *et al.* A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–123 (2008).
8. Avidor-Reiss, T. *et al.* Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* **117,** 527–539 (2004).
9. Enault, F., Suhre, K., Abergel, C., Poirot, O. & Claverie, J. M. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* **19** (Suppl. 1), i105–i107 (2003).
10. Drinnenberg, I. A. *et al.* RNAi in budding yeast. *Science* **326,** 544–550 (2009).
11. Drinnenberg, I. A., Fink, G. R. & Bartel, D. P. Compatibility with killer explains the rise of RNAi-deficient fungi. *Science* **333,** 1592 (2011).
12. Yelina, N. E. *et al.* Putative *Arabidopsis* THO/TREX mRNA export complex is involved in transgene and endogenous siRNA biosynthesis. *Proc. Natl Acad. Sci. USA* **107,** 13948–13953 (2010).
13. Ketting, R. F. & Plasterk, R. H. A genetic link between co-suppression and RNA interference in *C. elegans*. *Nature* **404,** 296–298 (2000).
14. Simmer, F. *et al.* Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr. Biol.* **12,** 1317–1319 (2002).
15. Cui, M., Kim, E. B. & Han, M. Diverse chromatin remodeling genes antagonize the Rb-involved SynMuv pathways in *C. elegans*. *PLoS Genet.* **2,** e74 (2006).
16. Parry, D. H., Xu, J. & Ruvkun, G. A whole-genome RNAi Screen for *C. elegans* miRNA pathway genes. *Curr. Biol.* **17,** 2013–2022 (2007).
17. Thivierge, C. *et al.* Tudor domain ERI-5 tethers an RNA-dependent RNA polymerase to DCR-1 to potentiate endo-RNAi. *Nature Struct. Mol. Biol.* **19,** 90–97 (2012).
18. Zhang, L. *et al.* Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell* **28,** 598–613 (2007).
19. Calvo, S. *et al.* Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature Genet.* **38,** 576–582 (2006).
20. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449–453 (2003).
21. Hibbs, M. A. *et al.* Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23,** 2692–2699 (2007).
22. Simonis, N. *et al.* Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nature Methods* **6,** 47–54 (2009).
23. Montgomery, T. A. *et al.* PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog henn-1. *PLoS Genet.* **8,** e1002616 (2012).
24. Bayne, E. H. *et al.* Splicing factors facilitate RNAi-directed silencing in fission yeast. *Science* **322,** 602–606 (2008).
25. Pontes, O. & Pikaard, C. S. siRNA and miRNA processing: new functions for Cajal bodies. *Curr. Opin. Genet. Dev.* **18,** 197–203 (2008).
26. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA* **97,** 11319–11324 (2000).
27. Bernstein, D. A. *et al.* *Candida albicans* Dicer (CaDcr1) is required for efficient ribosomal and spliceosomal RNA maturation. *Proc. Natl Acad. Sci. USA* **109,** 523–528 (2012).
28. Guang, S. *et al.* An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* **321,** 537–541 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Duchaine for access to his ERI-1 proteomic data before it was published and to S. Fischer, C. Zhang and T. Montgomery for helpful discussions. The work was supported by NIH GM088565 and the Pew Charitable Trusts (J.K.K.) and NIH GM44619 and GM098647 (G.R.).

Author Contributions Y.T., J.K.K. and G.R. designed experiments; Y.T. developed analytical tools and analysed data; and Y.T., A.C.B., G.D.H., M.A.N., S.M.G., H.G., R.K. and J.K.K. designed and carried out experiments. O.Z. gave statistical support and conceptual advice. Y.T., K.Y, B.C. and M.B. wrote code. Y.T., A.C.B., J.K.K. and G.R. wrote the paper. G.R. and J.K.K. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.R. (ruvkun@molbio.mgh.harvard.edu).

doi:10.1038/nature11791

The architecture of human general transcription factor TFIID core complex

Christoph Bieniossek[1]*, Gabor Papai[2]*, Christiane Schaffitzel[1], Frederic Garzoni[1], Maxime Chaillet[1], Elisabeth Scheer[3], Petros Papadopoulos[4], Laszlo Tora[3], Patrick Schultz[2] & Imre Berger[1]

The initiation of gene transcription by RNA polymerase II is regulated by a plethora of proteins in human cells. The first general transcription factor to bind gene promoters is transcription factor IID (TFIID). TFIID triggers pre-initiation complex formation, functions as a coactivator by interacting with transcriptional activators and reads epigenetic marks[1–3]. TFIID is a megadalton-sized multiprotein complex composed of TATA-box-binding protein (TBP) and 13 TBP-associated factors (TAFs)[3]. Despite its crucial role, the detailed architecture and assembly mechanism of TFIID remain elusive. Histone fold domains are prevalent in TAFs, and histone-like tetramer and octamer structures have been proposed in TFIID[4–6]. A functional core-TFIID subcomplex was revealed in *Drosophila* nuclei, consisting of a subset of TAFs (TAF4, TAF5, TAF6, TAF9 and TAF12)[7]. These core subunits are thought to be present in two copies in holo-TFIID, in contrast to TBP and other TAFs that are present in a single copy[8], conveying a transition from symmetry to asymmetry in the TFIID assembly pathway. Here we present the structure of human core-TFIID determined by cryo-electron microscopy at 11.6 Å resolution. Our structure reveals a two-fold symmetric, interlaced architecture, with pronounced protrusions, that accommodates all conserved structural features of the TAFs including the histone folds. We further demonstrate that binding of one TAF8–TAF10 complex breaks the original symmetry of core-TFIID. We propose that the resulting asymmetric structure serves as a functional scaffold to nucleate holo-TFIID assembly, by accreting one copy each of the remaining TAFs and TBP.

The overall shape of TFIID was unveiled by electron microscopy, revealing an asymmetric tri-lobed structure[9–12]. The paucity and heterogeneity of the endogenous material used limit structural insights to moderate resolution (~30 Å for human TFIID), prohibiting molecular level interpretation of TFIID architecture[3,11]. Endogenous yeast TFIID was analysed for subunit stoichiometry, revealing that a subset of six TAFs (TAF4, TAF5, TAF6, TAF9, TAF10 and TAF12) exist in two copies, whereas TBP and the remaining seven TAFs are present in a single copy[8]. The concept emerged in which TAFs present in duplicate form a two-fold symmetric scaffold, around which the remaining TAFs and TBP organize as peripheral subunits[7,12]. Studies in *Drosophila* cells revealed a functional core-TFIID complex, composed of TAF4, TAF5, TAF6, TAF9 and TAF12 *in vivo*[7]. In cryo-electron microscopy (cryo-EM) studies of yeast TFIID, a quasi-symmetric smaller shape was also found[13]. These results suggest the existence of a core-TFIID module of pivotal importance for the integrity and assembly of holo-TFIID[12].

We produced recombinant human core-TFIID complex, consisting of two copies each of TAF4, TAF5, TAF6, TAF9 and TAF12 (Supplementary Fig. 1). We determined the structure of this ~650 kilodalton (kDa) complex by single-particle cryo-EM (Fig. 1 and Supplementary Figs 2–5). The presence of two copies each of the five TAFs suggests a symmetric core-TFIID architecture, and a complete refinement without applying any symmetry constraint resulted in a structure exhibiting two-fold symmetry at 13.4 Å resolution (Supplementary Fig. 4). We refined the structure by imposing this symmetry constraint to a resolution of 11.6 Å (Supplementary Fig. 3).

The structure of human core-TFIID complex reveals an interlaced architecture and a remarkably large solvent accessible surface due to numerous protrusions and channels (Fig. 1). An iterative density truncation approach allowed us to place all conserved domains of the TAFs within core-TFIID (Supplementary Fig. 6). By fitting coordinates from crystal structures or homology models, and by biochemical engineering of key subunits, we could assign ~70% of the density to specific TAF domains (Supplementary Figs 6–10 and Supplementary Video 1).

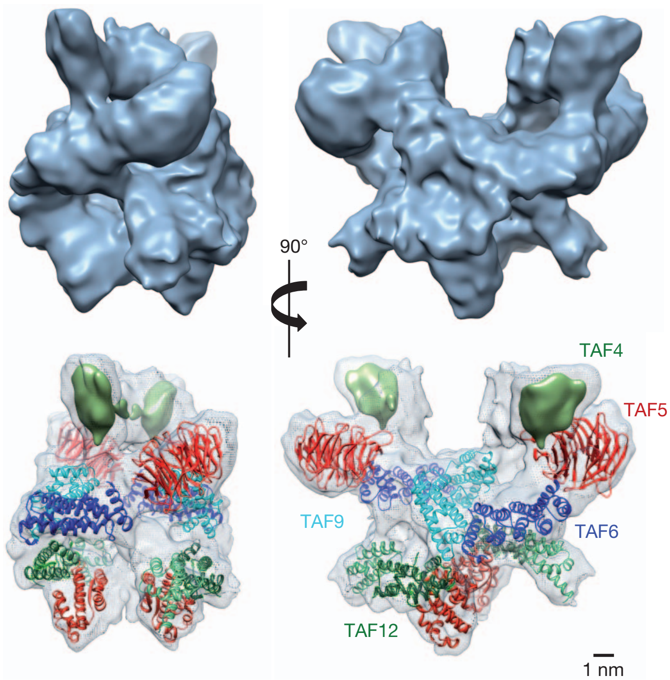


Figure 1 | Structure of the human TFIID core complex. The cryo-EM structure (top) is displayed in a side view (left) and from the front (right). The structural features in core-TFIID are shown (bottom). The cryo-EM density is transparent, TAF5 is coloured red (WD40 repeat domain, N-terminal domain), the TAF6 C-terminal domain is dark blue, and the TAF6 and TAF12 HF pair is light blue. The TAF4 N-terminal part, TAFH domain and HF pair with TAF12 are coloured green.

[1]European Molecular Biology Laboratory (EMBL) Grenoble Outstation, and Unit of Virus Host Cell Interactions UVHCI, UJF-CNRS-EMBL Unité Mixte International UMI 3265, 6 rue Jules Horowitz, 38042 Grenoble Cedex 9, France. [2]Department of Integrated Structural Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 1 rue Laurent Fries, BP10142, 67404 Illkirch, and U964 Inserm F-67400, and UMR7104 CNRS Illkirch, and Université de Strasbourg, F-67000 Strasbourg, France. [3]Department of Functional Genomics and Cancer, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 1 rue Laurent Fries, BP10142, 67404 Illkirch, and U964 Inserm F-67400, and UMR7104 CNRS Illkirch, and Université de Strasbourg, F-67000 Strasbourg, France. [4]Department of Cell Biology, Erasmus MC, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.
*These authors contributed equally to this work.

In the core-TFIID cryo-EM structure, a flat, slightly conical shape projecting from either side exhibits clear features of β-propeller structures characteristic of WD40 repeat domains. The TAF5 carboxy-terminal region contains six predicted WD40 repeats[14] (Fig. 2a). Our density shows six triangular knuckles consistent with six blades (Fig. 2b). TAF5 also contains a conserved amino-terminal domain (NTD) for which crystal structures exist[15,16]. Two protrusions at the bottom of the core-TFIID structure accommodate the crystal coordinates, suggesting that the TAF5 NTD is located distally from the C-terminal domain comprising the WD40 repeats (Fig. 2c). It has been proposed that the TAF5 NTD may have a role in the dimerization of TAF5 (ref. 15). In our structure, the TAF5 NTDs are not sufficiently close enough to engage an extended dimerization interface.

A recent crystal structure revealed a HEAT repeat domain in the TAF6 C-terminal part[17]. The TAF6 HEAT repeats are located adjacent to the TAF5 WD40 repeat domains, bracketing the front and back of the complex (Fig. 1). TAF6 also contains a conserved histone fold (HF) domain, which specifically interacts with a HF domain present in TAF9 to form an HF pair[4] (Fig. 2a). The cryo-EM density of core-TFIID exhibits four regions that can accommodate altogether two TAF6–TAF9 and two TAF4–TAF12 pairs. To assign their location, we determined the structure of a previously characterized ~400-kDa heterohexameric complex[18] containing two copies each of TAF5, TAF6 and TAF9 (hereafter denoted 3TAF) (Supplementary Figs 2 and 3). The 3TAF density reveals a holey basket-like structure with dimensions similar to core-TFIID, but lacking protrusions. The TAF5 WD40 repeat and NTD domains, the TAF6 HEAT repeats and the TAF6–TAF9 HF pairs are clearly discernible in the 3TAF structure, enabling unambiguous assignment of the TAF6–TAF9 HF pairs in core-TFIID (Supplementary Fig. 8). Note that we placed the pair as a unit as we cannot discriminate the TAF6 HF from the TAF9 HF. We also determined the cryo-EM structure of a mutant 3TAF complex containing TAF5 N-terminally tagged with maltose binding protein (MBP) to confirm the TAF5 NTD placement (Supplementary Fig. 9).

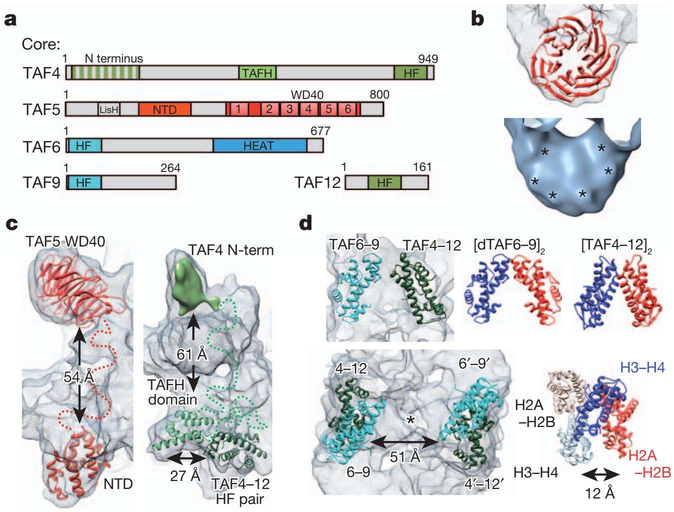


Figure 2 | Molecular organization of conserved TAF domains. a, TAF domain architecture (colour code as in Fig. 1). NTD denotes TAF5 N-terminal domain, WD40 denotes TAF5 C-terminal WD40 repeats. The TAF4 N-terminal part is hatched. HEAT denotes TAF6 C-terminal HEAT repeats; TAFH denotes conserved peptide-interaction domain in TAF4. LisH represents a non-conserved homology region. **b**, Density corresponding to the TAF5 WD40 repeat domain (top), with a closely related β-propeller (Protein Data Bank (PDB) accession 2PBI) superimposed (red). Six knuckles are marked by asterisks (bottom). **c**, Conformations adopted by TAF5 (red) and TAF4 (green) in core-TFIID. Amino acid stretches with unknown conformation are represented as dotted lines. **d**, Structural arrangement of TAF6–TAF9 and TAF4–TAF12 HF pairs, looking at the front (top, left). An identical arrangement is in the back. TAF6–TAF9 and TAF4–TAF12 tetramers present in crystals (top, right) and the histone octamer (bottom, right) are depicted for comparison. dTAF6–9, *Drosophila* TAF6–TAF9. Two copies each (marked by subscript '$_2$') of TAF6–TAF9, or TAF4–TAF12, respectively, form the tetramers in the crystals. The distance relating H2A–H2B and H3–H4 (12 Å) and the pseudo two-fold axis in the octamer (dashed line) are indicated. The HF pairs at the front (TAF4–12 and TAF6–9) and back (4′–12′ and 6′–9′) of the core-TFIID structure are separated by >50 Å (bottom, left). The core-TFIID two-fold axis is marked (asterisk).

TAF4 contains an N-terminal region of apparent low complexity, a central conserved domain called TAFH, and a conserved HF domain in the C-terminal region (Fig. 2a). The TAF4 HF domain pairs with TAF12 and atomic structures have been determined[19]. The difference density map between core-TFIID and 3TAF revealed the position of the two TAF4–TAF12 HF pairs, occupying density adjacent to the TAF6–TAF9 HF pairs. TAFH binds short hydrophobic peptides present in transcriptional regulators and the crystal structure shows a compact bundle of α-helices[20]. Two protrusions in the neighbourhood of the TAF4–TAF12 HF pairs accommodate the TAFH crystal coordinates (Fig. 2c). We prepared a mutant core-TFIID containing N-terminally truncated TAF4, and determined the electron microscopy structure of this complex (Supplementary Fig. 10). The two ear-like lobes on top of core-TFIID disappeared, indicating that this density corresponds to the TAF4 N-terminal parts. By contrast, the lateral protrusions in the lower part of the complex remained unaltered, confirming our TAFH placement.

The crystal structure of the *Drosophila* TAF6–TAF9 HF pair showed structural similarity with the heterotetrameric core of the histone octamer, formed by histones H3 and H4 (ref. 4) (Fig. 2d). Biochemical data suggested a similarity of TAF4 and TAF12 to histones H2A and H2B, respectively, leading to the proposal that a histone octamer-like structure may exist in TFIID[5,6]. Our cryo-EM structure of human core-TFIID contains two copies each of the TAF6–TAF9 and TAF4–TAF12 HF pairs. In the front and back of core-TFIID, one TAF6–TAF9 pair is juxtaposed to one TAF4–TAF12 pair (Fig. 2d). In contrast to the crystal structure that showed two identical dimers, the TAF6–TAF9 and TAF4–TAF12 HF pairs are less tightly associated and rotated with respect to each other. The distance across core-TFIID (>50 Å) rules out direct interactions between the two sets of HF pairs, whereas in the histone octamer, the H2A–H2B and H3–H4 pairs are within van der Waals contact. Our results suggest that the histone octamer-like arrangements mediated by TAF4, TAF6, TAF9 and TAF12 are not formed in TFIID.

Previous analyses of TAF locations relied on antibody mapping of yeast endogenous TFIID[21,22]. Our TAF5 geometry in core-TFIID is consistent with the immuno-mapping data, which detected two copies of TAF5, placed their N-terminal regions in close proximity, and mapped the C-terminal domains with the WD40 repeats to two opposite lobes in the holo-complex[21]. Likewise, the immuno-mapping study identified two copies of TAF6 and TAF9, in the vicinity of the TAF5 C-terminal parts. On the other hand, the TAF6–TAF9 and TAF4–TAF12 pairs arrange symmetrically in core-TFIID, whereas the immuno-mapping studies suggested an asymmetric arrangement[21]. This discrepancy may reflect errors in the immuno-mapping experiments, or, alternatively, may stem from changes in conformation or accessibility of the core-TFIID subunits, when further TAFs are accreted. Note that the immuno-mapping experiments used samples from yeast, whereas the present core-TFIID structure is from human. Yeast and human TAFs exhibit considerable variations in size, possibly affecting their geometries.

The structure of core-TFIID contains two copies each of its subunits in a symmetrical arrangement, whereas holo-TFIID, containing additional TAFs and TBP, has the shape of an asymmetric clamp[3]. How the structural transition from a symmetric to an asymmetric state occurs during TFIID assembly remains unknown. TAF8 regulates the nuclear import of TAF10, and both TAFs were shown to be co-imported as a complex into the nucleus by an importin α/β-dependent pathway[23]. Combinatorial assembly experiments showed that the TAF8–TAF10

pair can only be incorporated into a larger complex when all five TAFs forming core-TFIID are present[24]. Thus, we proposed that the transition from a symmetric core-TFIID to an asymmetric assembly may occur at the step of TAF8–TAF10 complex integration and may be regulated by nuclear-import mechanisms. We prepared a complex comprising core-TFIID and TAF8–TAF10, and determined the structure of this ~710-kDa complex (hereafter termed 7TAF) by cryo-EM (Fig. 3).

The 7TAF complex structure shows major perturbations when compared to core-TFIID, and notably deviates from two-fold symmetry (Fig. 3). Careful inspection of the cryo-EM density reveals that two different parts can be defined in the 7TAF structure. One half adopts largely the same shape as in core-TFIID, but most of the rearrangements localize to the other half (Fig. 3a and Supplementary Fig. 11). The ear-like lobe on top swings over by ~40 Å in this rearranged half, and new density is present at the bottom in the vicinity of the TAF5 NTD. We reasoned that this new density could be attributed to the binding of TAF8–TAF10. We used a mouse monoclonal antibody (mAb6TA) that specifically binds TAF10 to prepare a 7TAF–mAb6TA complex. Electron microscopy analysis revealed binding of the antibody to the bottom part of the structure, confirming the position of TAF8–TAF10 (Supplementary Fig. 12). The new density is consistent with the volume occupied by one HF pair, suggesting that a single TAF8–TAF10 heterodimer is incorporated into the 7TAF complex. Reconstitution experiments of core-TFIID and wild-type TAF8–TAF10 complex at defined ratios were consistent with the presence of one copy of TAF8–TAF10 in the 7TAF complex. Experiments involving a MBP-tagged variant of TAF8 confirmed this finding (Supplementary Fig. 13). The new density in the 7TAF complex extends over the two-fold axis that previously related the two halves of core-TFIID (Fig. 3b and Supplementary Fig. 12). Steric hindrance thus rules out the incorporation of a second TAF8–TAF10 copy. We demonstrated that the stoichiometry of TAFs in our recombinant complexes, notably the existence of a single copy of TAF8, is the same as in endogenous human TFIID by comparative western blots, and by protein abundance determination following mass spectrometry analyses (Supplementary Figs 14 and 15). Our results provide a mechanistic model for the structural transition of TFIID from a symmetric core to the asymmetric holo-complex (Fig. 4).

The TAF6–TAF9 pair has been shown to bind the downstream core promoter element in TATA-less promoters specifically[25]. Our structure places the TAF6–TAF9 pairs to the surface, well positioned to interact with downstream core promoter elements. TAF5, TAF6 and TAF9 were identified in both TFIID and the large coactivator SAGA[26] in yeast. In humans, gene duplication resulted in TAF5L and TAF6L, which are closely related variants substituting for TAF5 and TAF6 in human SAGA[26]. We propose that the 3TAF structure constitutes the common central scaffold of TFIID and SAGA. TAF variants mediating specific cellular functions have been identified[27–30]. In cells in which TAF4b is expressed, TAF4 and TAF4b co-exist in TFIID. TAF4b contains a different N-terminal region than TAF4, thus the structural basis of the open conformation found for TAF4b-containing TFIID complexes may reside in distinct geometries of the ear-like lobes[28]. TAF6δ was found to link apoptotic signalling pathways to TFIID function[29]. TAF6δ has a deletion in its HF domain, and TAF6δ-containing TFIID lacks the HF partner of TAF6, TAF9 (ref. 29). Our 3TAF structure shows numerous interfaces between TAF5, TAF6 and TAF9, and the loss of TAF9 due to the compromised HF in TAF6δ may be tolerated to some extent during TFIID assembly. The TAF9 related factor, TAF9b, was implicated in gene silencing and transcriptional repression[30]. TAF9b and TAF9 have very similar sequences, and we expect that incorporation of TAF9b will not cause major rearrangements.

We determined the structures of three distinct TFIID subassemblies, providing a molecular framework for rationalizing TFIID core architecture. Conformational changes are found between the structures, most pronounced when 7TAF is formed from core-TFIID and TAF8–TAF10. Our structures suggest that the step-wise assembly of partial TFIID complexes recapitulates molecular events along a pathway leading to holo-TFIID in cells.

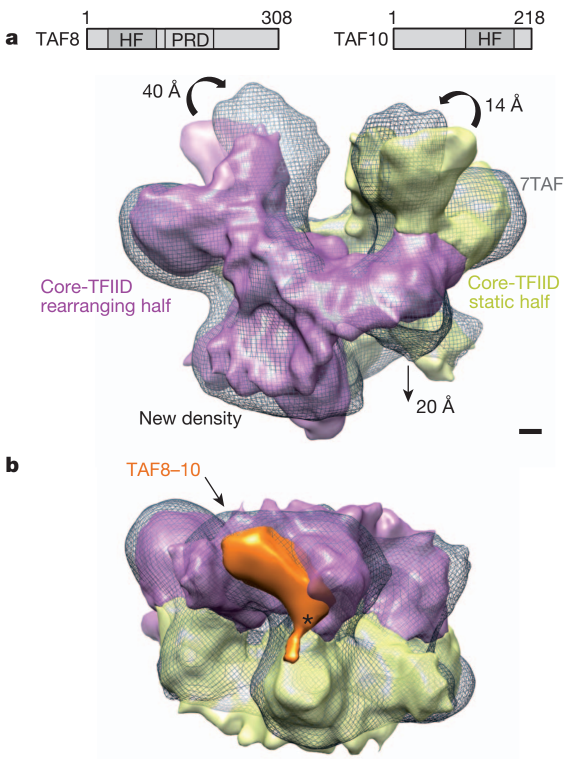


Figure 3 | Asymmetric 7TAF structure. **a**, TAF8 and TAF10 are represented as bars (top) showing HFs and a proline-rich domain (PRD). The 7TAF structure (grey mesh) is superimposed on core-TFIID coloured in purple (rearranging half) and yellow (static half). Conformational changes are marked with arrows. New density is observed in 7TAF. Scale bar, 1 nm. **b**, 7TAF (mesh) in a bottom view, superimposed on core-TFIID. New density in the 7TAF structure is drawn in orange. The two-fold axis of core-TFIID is marked (asterisk).

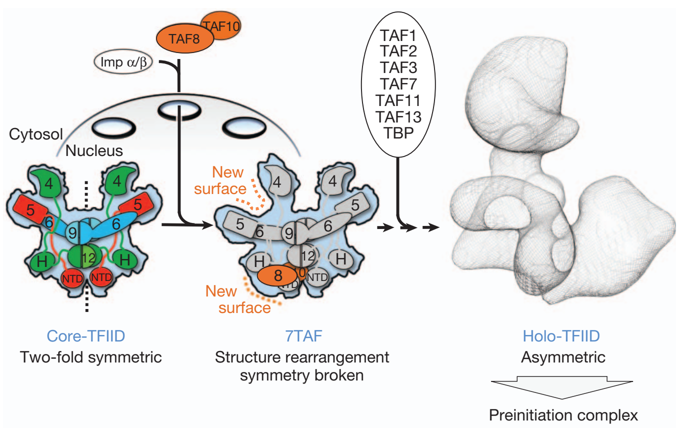


Figure 4 | Model for holo-TFIID assembly. Core-TFIID with two copies of TAF4, TAF5, TAF6, TAF9 and TAF12 is symmetric (left). The TAF8–TAF10 complex (orange) is imported into the nucleus by importins (imp)[23] (top). Binding of one copy of TAF8–TAF10 breaks the symmetry in core-TFIID, resulting in an asymmetric 7TAF complex (middle). 7TAF exhibits two distinct halves and new binding surfaces for further subunits (dashed lines). Accretion of remaining TAFs and TBP in single copy, results in asymmetric clamp-shaped holo-TFIID (EMD-1195, grey mesh) that nucleates the preinitiation complex (right).

METHODS SUMMARY

Recombinant TAF complexes were produced using MultiBac and polyproteins. Endogenous TFIID for mass spectrometry was purified from cultured HeLa or fetal liver cells. Negative-stain electron microscopy was performed and two-dimensional class averages were calculated (IMAGIC software package) as a benchmark for optimizing complex purification protocols until satisfactory sample quality was achieved. For cryo-EM grid preparation, TAF complexes were stabilized by mild glutaraldehyde cross-linking in a glycerol gradient. The specimens were adsorbed on a thin carbon film sustained by a holey carbon grid and plunge-frozen in liquid ethane with controlled temperature and humidity. Images of the 3TAF and core-TFIID complexes were recorded at 50,000× on a cryo-transmission electron microscope (cryo-TEM, Tecnai F20) at 200 kV, digitized on a drum scanner (Primescan D7100) at 5,000 dpi. 3TAF was coarsened twice resulting in a pixel spacing of 2.03 Å, core-TFIID was coarsened three times resulting in a pixel spacing of 3.05 Å before analysis on the specimen. CCD frames of the 7TAF complex were recorded at 58,000× on a cryo-TEM (Tecnai Polara) operating at 100 kV and coarsened twice resulting in a final pixel spacing of 3.72 Å. Images were selected using the EMAN2 software package and analysed with the IMAGIC, Bsoft and Spider software packages. The resolution of the structures was determined according to the 0.5 cut-off of the Fourier Shell Correlation curve and the final reconstructions were filtered accordingly. Interactive fitting of atomic structures and generation of images for publication were performed using the University of California, San Francisco, Chimera software.

Received 31 March; accepted 14 November 2012.
Published online 6 January 2013.

1. Papai, G. *et al.* TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation. *Nature* **465,** 956–960 (2010).
2. Müller, F., Zaucker, A. & Tora, L. Developmental regulation of transcription initiation: more than just changing the actors. *Curr. Opin. Genet. Dev.* **20,** 533–540 (2010).
3. Papai, G., Weil, P. A. & Schultz, P. New insights into the function of transcription factor TFIID from recent structural studies. *Curr. Opin. Genet. Dev.* **21,** 219–224 (2011).
4. Xie, X. *et al.* Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature* **380,** 316–322 (1996).
5. Hoffmann, A. *et al.* A histone octamer-like structure within TFIID. *Nature* **380,** 356–359 (1996).
6. Selleck, W. *et al.* A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. *Nature Struct. Biol.* **8,** 695–700 (2001).
7. Wright, K. J., Marr, M. T. II & Tjian, R. TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. *Proc. Natl Acad. Sci. USA* **103,** 12347–12352 (2006).
8. Sanders, S. L., Garbett, K. A. & Weil, P. A. Molecular characterization of Saccharomyces cerevisiae TFIID. *Mol. Cell. Biol.* **22,** 6000–6013 (2002).
9. Andel, F. III, Ladurner, A. G., Inouye, C., Tjian, R. & Nogales, E. Three-dimensional structure of the human TFIID-IIA-IIB complex. *Science* **286,** 2153–2156 (1999).
10. Brand, M., Leurent, C., Mallouh, V., Tora, L. & Schultz, P. Three-dimensional structures of the TAFII-containing complexes TFIID and TFTC. *Science* **286,** 2151–2153 (1999).
11. Grob, P. *et al.* Cryo-electron microscopy studies of human TFIID: conformational breathing in the integration of gene regulatory cues. *Structure* **14,** 511–520 (2006).
12. Cler, E., Papai, G., Schultz, P. & Davidson, I. Recent advances in understanding the structure and function of general transcription factor TFIID. *Cell. Mol. Life Sci.* **66,** 2123–2134 (2009).
13. Papai, G. *et al.* Mapping the initiator binding Taf2 subunit in the structure of hydrated yeast TFIID. *Structure* **17,** 363–373 (2009).
14. Dubrovskaya, V. *et al.* Distinct domains of hTAFII100 are required for functional interaction with transcription factor TFIIF beta (RAP30) and incorporation into the TFIID complex. *EMBO J.* **15,** 3702–3712 (1996).
15. Bhattacharya, S., Takada, S. & Jacobson, R. H. Structural analysis and dimerization potential of the human TAF5 subunit of TFIID. *Proc. Natl Acad. Sci. USA* **104,** 1189–1194 (2007).
16. Romier, C. *et al.* Crystal structure, biochemical and genetic characterization of yeast and E. cuniculi TAF(II)5 N-terminal domain: implications for TFIID assembly. *J. Mol. Biol.* **368,** 1292–1306 (2007).
17. Scheer, E., Delbac, F., Tora, L., Moras, D. & Romier, C. TFIID TAF6–TAF9 complex formation involves the HEAT repeat-containing C-terminal domain of TAF6 and is modulated by TAF5. *J. Biol. Chem.* **287,** 27580–27592 (2012).
18. Fitzgerald, D. J. *et al.* Multiprotein expression strategy for structural biology of eukaryotic complexes. *Structure* **15,** 275–279 (2007).
19. Werten, S. *et al.* Crystal structure of a subcomplex of human transcription factor TFIID formed by TATA binding protein-associated factors hTAF4 ($hTAF_{II}135$) and hTAF12 ($hTAF_{II}20$). *J. Biol. Chem.* **277,** 45502–45509 (2002).
20. Wang, X. *et al.* Conserved region I of human coactivator TAF4 binds to a short hydrophobic motif present in transcriptional regulators. *Proc. Natl Acad. Sci. USA* **104,** 7839–7844 (2007).
21. Leurent, C. *et al.* Mapping histone fold TAFs within yeast TFIID. *EMBO J.* **21,** 3424–3433 (2002).
22. Leurent, C. *et al.* Mapping key functional sites within yeast TFIID. *EMBO J.* **23,** 719–727 (2004).
23. Soutoglou, E. *et al.* The nuclear import of TAF10 is regulated by one of its three histone fold domain-containing interaction partners. *Mol. Cell. Biol.* **25,** 4092–4104 (2005).
24. Demény, M. A. *et al.* Identification of a small TAF complex and its role in the assembly of TAF-containing complexes. *PLoS ONE* **2,** e316 (2007).
25. Burke, T. W. & Kadonaga, J. T. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.* **11,** 3020–3031 (1997).
26. Timmers, H. T. M. & Tora, L. SAGA unveiled. *Trends Biochem. Sci.* **30,** 7–10 (2005).
27. Mengus, G. *et al.* TAF4 inactivation in embryonic fibroblasts activates TGFβ signaling and autocrine growth. *EMBO J.* **24,** 2753–2767 (2005).
28. Liu, W.-L. *et al.* Structural changes in TAF4b–TFIID correlate with promoter selectivity. *Mol. Cell* **29,** 81–91 (2008).
29. Bell, B., Scheer, E. & Tora, L. Identification of hTAFII80δ links apoptotic signaling pathways to transcription factor TFIID function. *Mol. Cell* **8,** 591–600 (2001).
30. Chen, Z. & Manley, J. L. *In vivo* functional analysis of the histone 3-like TAF9 and a TAF9-related factor, TAF9L. *J. Biol. Chem.* **278,** 35172–35183 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all members of the Berger, Schultz and Tora laboratories for advice and discussions. We are grateful to C. Romier for providing the X-ray structure of the TAF6 C-terminal domain before publication. We thank J. Demmers for mass spectrometric analyses and F. Grosveld for discussions. The EMBL, IBS and IGBMC core facilities are acknowledged for services. We are indebted to G. Schoehn for maintaining the electron microscopes in Grenoble. T. J. Richmond is acknowledged for advice and support. C.B. is a fellow of the joint European Commission (EC)/EMBL interdisciplinary research opportunities program (EIPOD). C.S. is recipient of a European Research Council (ERC) Starting Grant and an Agence Nationale de la Recherche (ANR) Jeunes Chercheuses award. I.B. acknowledges support from the EC Marie Curie Action and the EC Framework Programme (FP) 7 projects INSTRUCT, PCUBE, BioSTRUCT-X, 4D-CellFate and ComplexINC. P.S. acknowledges support from the Institut National de la Santé et de la Recherche Médicale (INSERM), the Centre National pour la Recherche Scientifique (CNRS), the Association pour la Recherche sur le Cancer (ARC) and the Fondation pour la Recherche Médicale (FRM). This work was supported by the ANR Projets Blancs puzzle-fit (to P.S.), ChromAct (to P.S. and L.T.) and TFIID-Complexes (to L.T., P.S. and I.B.).

Author Contributions P.S., L.T. and I.B. designed the study; C.B., F.G. and I.B. implemented the MultiBac system; C.B., F.G. and M.C. produced, purified and characterized all TAF complexes; C.S. implemented gradient centrifugation and GraFix, analysed negative-stain electron microscopy data and calculated two-dimensional class averages; G.P. and P.S. carried out random conical tilt experiments, collected and analysed cryo-EM data of all complexes, and calculated and refined the electron microscopy densities; G.P. prepared the core-TFIID molecular structure by fitting crystal coordinates and homology models; E.S., L.T. and P.P. prepared and analysed endogenous TFIID for protein content. L.T. provided the anti-TAF10 antibody (mAb6TA). P.S., L.T. and I.B. supervised the work. G.P., C.B., P.S., L.T. and I.B. prepared the figures and wrote the manuscript together.

Author Information The cryo-EM maps have been deposited in the 3D-EM database (EMBL-European Bioinformatics Institute). EMBD accession codes are EMD-2229 (3TAF), EMD-2230 (core-TFIID) and EMD-2231 (7TAF). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.S. (patrick.schultz@igbmc.fr) or I.B. (iberger@embl.fr).

doi:10.1038/nature11778

Rotation mechanism of *Enterococcus hirae* V_1-ATPase based on asymmetric crystal structures

Satoshi Arai[1,2]*, Shinya Saijo[2,3]*, Kano Suzuki[1]*, Kenji Mizutani[1,2,4], Yoshimi Kakinuma[5], Yoshiko Ishizuka-Katsura[6], Noboru Ohsawa[6], Takaho Terada[6], Mikako Shirouzu[6], Shigeyuki Yokoyama[6,7,8], So Iwata[4,6], Ichiro Yamato[2] & Takeshi Murata[1,6,9]

In various cellular membrane systems, vacuolar ATPases (V-ATPases) function as proton pumps, which are involved in many processes such as bone resorption and cancer metastasis, and these membrane proteins represent attractive drug targets for osteoporosis and cancer[1]. The hydrophilic V_1 portion is known as a rotary motor, in which a central axis DF complex rotates inside a hexagonally arranged catalytic A_3B_3 complex using ATP hydrolysis energy, but the molecular mechanism is not well defined owing to a lack of high-resolution structural information. We previously reported on the *in vitro* expression, purification and reconstitution of *Enterococcus hirae* V_1-ATPase from the A_3B_3 and DF complexes[2,3]. Here we report the asymmetric structures of the nucleotide-free (2.8 Å) and nucleotide-bound (3.4 Å) A_3B_3 complex that demonstrate conformational changes induced by nucleotide binding, suggesting a binding order in the right-handed rotational orientation in a cooperative manner. The crystal structures of the nucleotide-free (2.2 Å) and nucleotide-bound (2.7 Å) V_1-ATPase are also reported. The more tightly packed nucleotide-binding site seems to be induced by DF binding, and ATP hydrolysis seems to be stimulated by the approach of a conserved arginine residue. To our knowledge, these asymmetric structures represent the first high-resolution view of the rotational mechanism of V_1-ATPase.

V-ATPases are thought to have originated from an ancestral enzyme in common with F-ATPases, which function as ATP synthases in mitochondria, chloroplasts and oxidative bacteria[4,5]. These ATPases possess an overall similar structure that is composed of a hydrophilic domain (V_1 and F_1) and a membrane-embedded ion-transporting domain (V_o and F_o), and they have a similar reaction mechanism that occurs through rotation[1]. The rotational catalysis of F_1-ATPase has been investigated in detail, and the molecular mechanism has been proposed on the basis of crystal structures of the complex from bovine[6–9], yeast[10–12] and bacteria[13,14], and extensive single-molecule observation of the rotation[15–17]. Similar V_1-ATPase experiments have been conducted using the *Thermus thermophilus* enzyme, which functions physiologically as an ATP synthase[18]. The crystal structures of the A_3B_3 complex at 2.8 Å resolution[19] and the A_3B_3DF (V_1) complex at low resolution (4.5–4.8 Å)[20] suggest differences in its structure and interactions compared to F_1-ATPases. Single-molecule analyses of V_1-ATPase also suggest differences in torque generation and the coupling scheme of the rotation mechanism as compared to F_1 (ref. 21).

Enterococcus hirae V-ATPase, which acts as a primary ion pump similar to eukaryotic V-ATPases, uniquely transports Na^+ or Li^+ instead of H^+ ions[22–25]. The enzyme is composed of nine subunits with amino acid sequences that are homologous to those of the corresponding subunits of eukaryotic V-ATPases[26–28] (Supplementary Fig. 1). In this study, we solved the first asymmetric structures of A_3B_3 and A_3B_3DF (V_1) complexes at high resolution, which enabled the generation of a new model of the rotational mechanism.

The *E. hirae* A_3B_3 complex was purified and crystallized in the absence of nucleotide. The crystal structure was solved at a resolution of 2.8 Å (Supplementary Table 1). The three catalytic A subunits (Eh-A) and the three non-catalytic B subunits (Eh-B) are alternatively arranged and form a hexagonal ring (Fig. 1a). The structures of these subunits comprise the amino-terminal β-barrel domain, the central α/β domain and the carboxy-terminal helical domain (Supplementary Figs 2 and 3). The three Eh-A or Eh-B subunits in the A_3B_3 complex were shown to have similar secondary structures, but their three-dimensional conformations slightly differed (see Supplementary Fig. 4). We superimposed the N-terminal β-barrel region of the three Eh-A or Eh-B subunits to examine the conformational differences in the A_3B_3 complex, because this β-barrel domain should be fixed to form an alternatively arranged ring. One of the three Eh-A subunits adopts a closed conformation (denoted as A_C), which shifts the structure into the centre of the A_3B_3 ring, whereas the other two Eh-A subunits adopt similar open conformations (denoted as A_O and $A_{O'}$), even though an α-helix (residues 261–275; designated as the 'arm') of the three Eh-A conformations was almost fixed (Fig. 1b and Supplementary Fig. 4). Similarly, one of the three Eh-B subunits shows a closed conformation (denoted as B_C) compared to the others (denoted as B_O and $B_{O'}$) (Fig. 1c). Thus, the A_3B_3 hexamer assembled asymmetrically by adjacent A_O and B_O, $A_{O'}$ and $B_{O'}$ and A_C and B_C subunits, whereas the conserved nucleotide-binding sites were located between the three different combinations: A_OB_C, $A_{O'}B_O$ and $A_CB_{O'}$ pairs (Fig. 1d).

The nucleotide-binding sites, which are comprised of the phosphate-binding loop (P-loop: GXXXXGKT(S)), the N-terminal portion of the arm (Glu 261 and Arg 262) in Eh-A, and Arg 350 (the 'Arg-finger' in ATPases) in Eh-B, had three different conformations (Fig. 1i–k). Electron density for nucleotides in the three binding sites was not observed (Supplementary Fig. 5), consistent with the absence of nucleotide contamination (ADP and ATP) detected in our sample and nucleotide-free crystallization conditions. Thus, surprisingly, the A_3B_3 complex, which is formed by three identical Eh-A and Eh-B subunits, demonstrated complete asymmetry without nucleotide binding (see Supplementary Fig. 6 for comparison with the symmetric structures of corresponding previously reported complexes of V- and F-ATPases). The observed asymmetric structure of nucleotide-free A_3B_3 (designated as eA_3B_3) may provide new insights for understanding the cooperative nature of the rotary motor as described below.

Next, we crystallized the A_3B_3 complex in the presence of a high concentration (5 mM) of the non-hydrolysable ATP analogue adenosine 5′-(β,γ-imino)triphosphate (AMP-PNP) with $MgSO_4$, and solved the structure of nucleotide-bound A_3B_3 (denoted as bA_3B_3) at

[1]Department of Chemistry, Graduate School of Science, Chiba University, 1-33 Yayoi-cho, Inage, Chiba 263-8522, Japan. [2]Department of Biological Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba 278-8510, Japan. [3]RIKEN SPring-8 Center, 1-1-1 Kouto, Sayo, Hyogo 679-5148, Japan. [4]Department of Cell Biology, Faculty of Medicine, Kyoto University, Yoshidakonoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. [5]Laboratory of Molecular Physiology and Genetics, Faculty of Agriculture, Ehime University, 3-5-7 Tarumi, Matsuyama, Ehime 790-8566, Japan. [6]RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan. [7]Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [8]Laboratory of Structural Biology, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [9]JST, PRESTO, 1-33 Yayoi-cho, Inage, Chiba 263-8522, Japan.
*These authors contributed equally to this work.

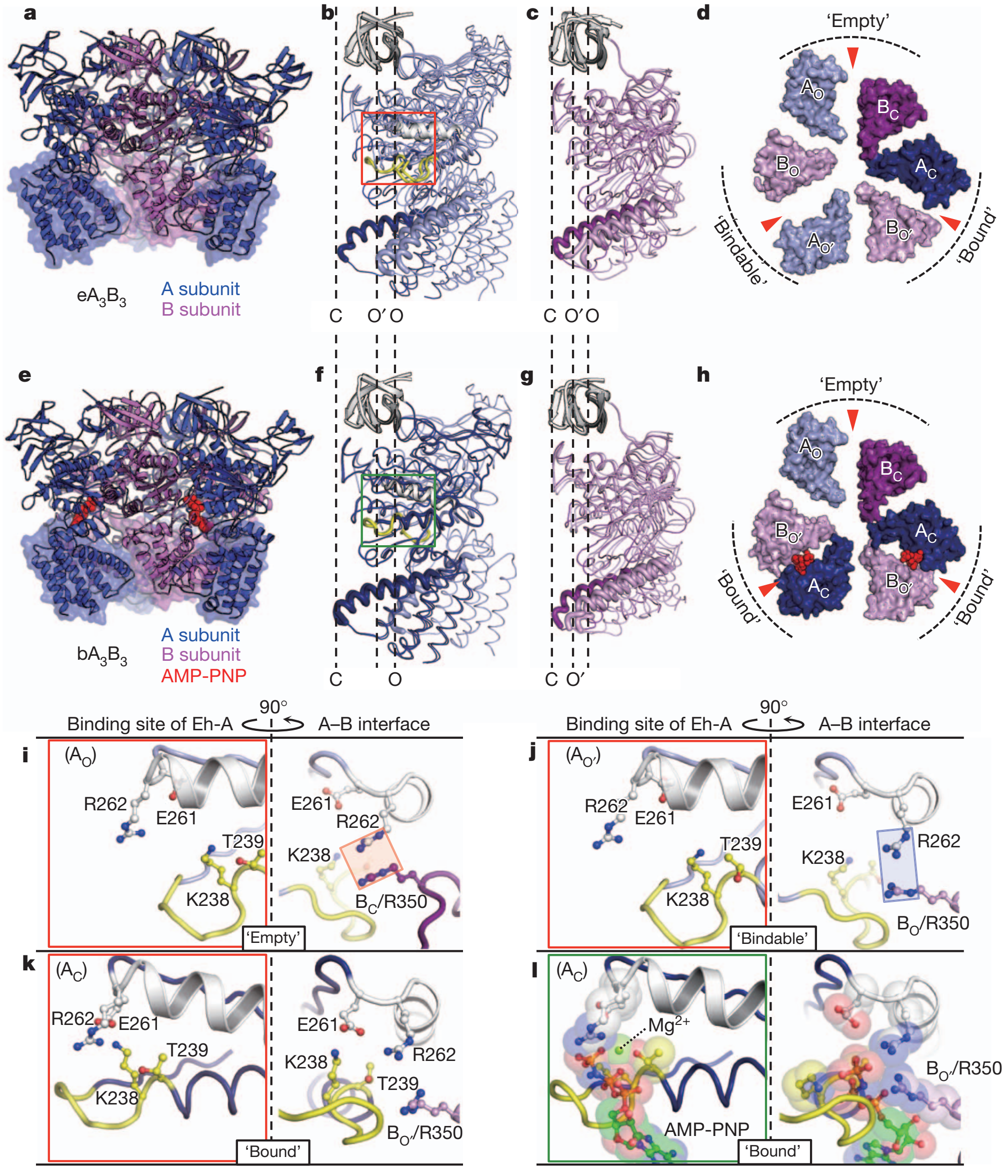


Figure 1 | Structure of the A$_3$B$_3$ complex. **a**, Side view of the nucleotide-free A$_3$B$_3$ structure (eA$_3$B$_3$). **b**, **c**, Superimposed structures at the N-terminal β-barrel (white) of three structures of Eh-A (**b**) and Eh-B (**c**) in eA$_3$B$_3$. Open (O and O') and closed (C) conformations of Eh-A and Eh-B are shown in light and dark colours, respectively. The P-loop and arm are shown in yellow and white, respectively. **d**, Top view of the C-terminal domain (shown in **a** as transparent surface) of eA$_3$B$_3$ from the N-terminal β-barrel side. Red arrows indicate the nucleotide-binding sites. **e**–**h**, Structures of the AMP-PNP-bound A$_3$B$_3$ complex (bA$_3$B$_3$) viewed and coloured as in **a**–**d**. **i**–**l**, Magnified nucleotide-binding sites with conserved residues, corresponding to red (**b**) and green (**f**) boxes. Right panels show the A–B interfaces rotated 90° around a vertical axis from the left panels.

a resolution of 3.4 Å (Supplementary Table 1). Two strong electron density peaks for AMP-PNP:Mg were found at the nucleotide-binding pockets in two Eh-AB pairs (Supplementary Fig. 7). The other AB pair, in which no density for nucleotide was found, was very similar to the A$_O$B$_C$ pair in eA$_3$B$_3$ (root mean squared deviation (r.m.s.d.) = 0.477 Å). We designated these A$_O$B$_C$ pairs as the 'empty' form on the basis of their apparent very low affinity for AMP-PNP:Mg. The two AMP-PNP:Mg-bound AB pairs were very similar to each other (r.m.s.d. = 0.511 Å), and were also similar to the A$_C$B$_{O'}$ pair in eA$_3$B$_3$ (r.m.s.d. = 0.683 Å and 0.719 Å) except for the side-chain conformations that directly interacted with AMP-PNP:Mg (Fig. 1k, l and Supplementary Fig. 8). This finding suggests that the A$_C$B$_{O'}$ pair of eA$_3$B$_3$ takes the ATP-bound form even in the absence of nucleotide; we designated these A$_C$B$_{O'}$ pairs as the 'bound' form. Furthermore, the more open A$_{O'}$B$_O$ pair of eA$_3$B$_3$ seemed to bind AMP-PNP:Mg and to change to the bound form of bA$_3$B$_3$; that is, binding of AMP-PNP:Mg induced the conformational change of A$_3$B$_3$ from the A$_{O'}$B$_O$ pair to the A$_C$B$_{O'}$ pair (see Supplementary Video 1). We designated this unique A$_{O'}$B$_O$ pair of eA$_3$B$_3$ as 'bindable' form.

The reason that the empty form cannot bind AMP-PNP:Mg whereas the bindable form can is discussed here. The binding sites of empty and bindable forms are very similar except for the topologies of the Arg-fingers (Arg 350): the Arg-finger of closed Eh-B (B$_C$) in the empty conformation was closer to Arg 262 than that of open Eh-B (B$_O$) in the bindable conformation (Fig. 1i, j, red and blue boxes), which may prevent AMP-PNP:Mg binding. The conformation of Eh-B may regulate ATP-binding affinity by the Arg-finger (Arg 350) at the binding sites. Thus, these asymmetric structures suggest that the formation of the A$_3$B$_3$ hexamer ring imposes a restriction (stress) on the Eh-AB pair to induce conformational changes (strains) that cooperatively generate one empty (ATP-unbound form), one bindable (ATP-accessible form) and one bound (ATP-bound form) conformation, which in turn determines the order of nucleotide binding in the ring in the right-handed rotational orientation viewed from the top of the V$_1$ complex.

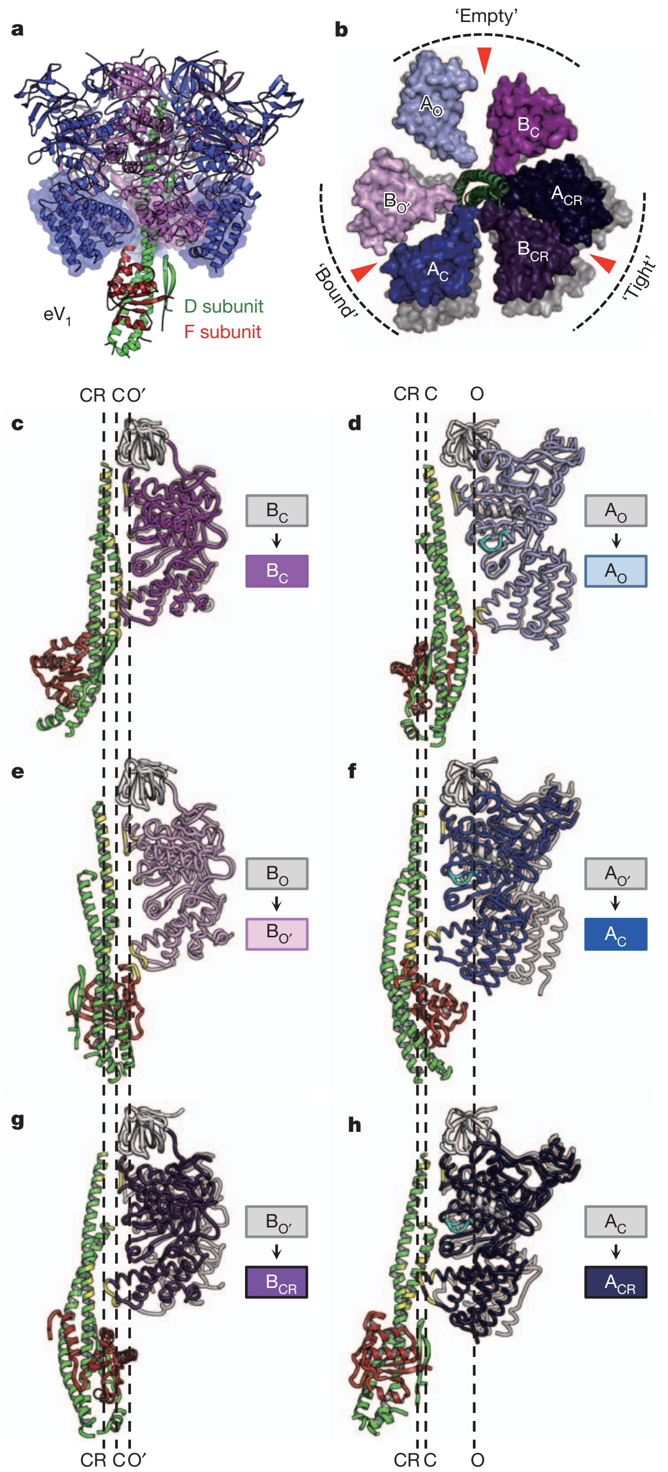


Figure 2 | Comparison of the asymmetric structures of nucleotide-free A_3B_3DF and A_3B_3 complexes. **a**, Side view of the nucleotide-free A_3B_3DF structure (eV_1). **b**, Top views of the C-terminal domain of eV_1 as in Fig. 1d, which is superimposed at the empty form onto that of transparent eA_3B_3 (grey). Open (O and O'), closed (C) and closer (CR) conformations of Eh-A and -B are shown in light, dark and darker colours, respectively. **c**–**h**, Protein–protein interactions between A_3B_3 and DF in eV_1. The B_C (**c**), A_O (**d**), $B_{O'}$ (**e**), A_C (**f**), B_{CR} (**g**) and A_{CR} (**h**) with DF complex in eV_1 are shown in side-viewed ribbon representation, which are compared with corresponding subunits (grey) of eA_3B_3 superimposed as in **b**. The P-loop is shown in cyan. The residues with buried surface area >10 Å^2, as calculated by PDBePISA (http://pdbe.org/pisa/), are shown in yellow.

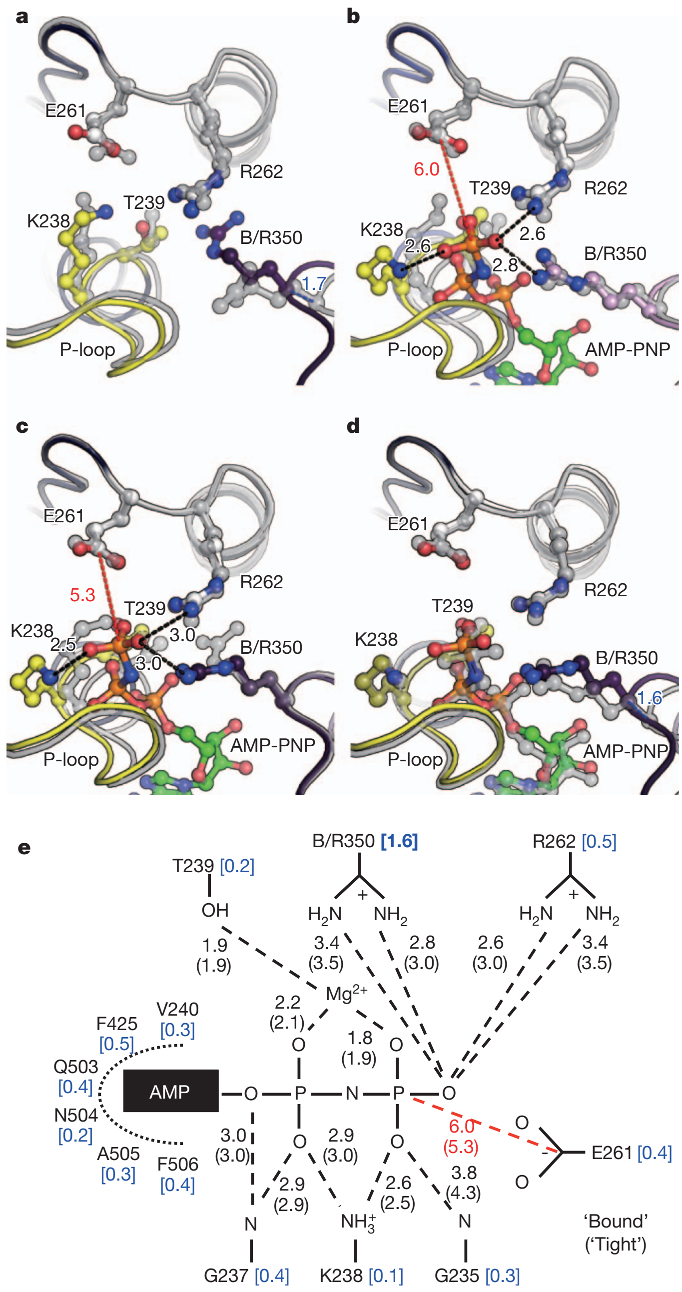


Figure 3 | Comparison of the nucleotide-binding sites. **a**–**d**, The viewing position, colours and representations of the binding site correspond to those of the right columns in Fig. 1i–l. These structures were superimposed at Eh-A (residues 67–593) of the compared AB pairs. **a**, Tight form in eV_1 (colour) compared with bound form in eA_3B_3 (grey). **b**, Bound form in bV_1 (colour) compared with bound form in eV_1 (grey). **c**, Tight form in bV_1 (colour) compared with tight form in eV_1 (grey). **d**, Tight form in bV_1 (colour) compared with bound form in bV_1 (grey). The distances (Å) between atoms are shown with dotted lines. **e**, A schematic representation of the nucleotide-binding sites of bV_1. The distances (Å) between atoms in the bound form or tight form (shown in parentheses) are shown with dotted lines. The distances (Å) between Cαs in the superimposed structure (**d**) are shown in blue brackets.

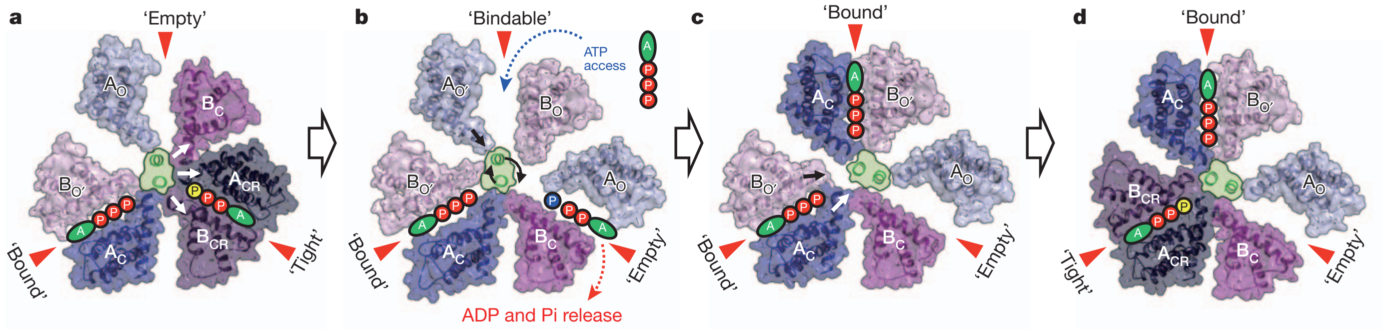


Figure 4 | A model of the rotation mechanism of V_1-ATPase. a–d, The structure models are on the basis of the crystal structures of bV_1 (**a** and **d**), eA_3B_3 (**b**) and bA_3B_3 (**c**) in this study. ATP with yellow 'P' in **a** and **d** represents an ATP molecule that is committed to hydrolysis. The blue 'P' in **b** represents a Pi molecule after hydrolysis of ATP. See text for further details.

Next, we crystallized and solved the crystal structure of the nucleotide-free V_1-ATPase (denoted as eV_1) at 2.2 Å resolution (see Supplementary Figs 9–11 and Supplementary Discussion for details). Eh-A and Eh-B assembled asymmetrically, similar to the A_3B_3 complex, and a central axis composed of Eh-D and Eh-F penetrated into the cavity of the A_3B_3 hexamer (Fig. 2a). Eh-D demonstrated a straighter conformation compared with that of the crystal structure of the DF complex[3], but other features were very similar (Supplementary Fig. 12). The coiled-coil α-helices of Eh-D interacted with several residues inside the A_3B_3 complex by forming 19 polar interactions and 101 non-polar (van der Waals) interactions (Fig. 2c–h and Supplementary Fig. 13).

The structural differences between eA_3B_3 and eV_1 that should have been induced by interaction with the DF complex are compared in Fig. 2. The eV_1 had an empty form (A_OB_C) (r.m.s.d = 0.544 Å) and a bound form ($A_CB_{O'}$) (r.m.s.d. = 0.699 Å), but the eV_1-bound form was positioned as the site of the eA_3B_3-bindable form when both empty forms were superimposed (Figs 1d and 2b and Supplementary Fig. 14). Therefore, the DF binding seemed to induce a change from the bindable form ($A_{O'}B_O$) of eA_3B_3 to the bound form ($A_CB_{O'}$), similar to the conformational changes of the eA_3B_3 induced by AMP-PNP binding (see Supplementary Videos 1 and 2). The remaining AB pair of eV_1 represented a more tightly packed conformation composed of closer Eh-A and -B conformations approaching the centre of the A_3B_3 ring, and this was not observed in the structure of the A_3B_3 complex (Fig. 2g, h). We designated these new conformations of closer Eh-A and -B subunits and the tightly packed Eh-AB pair as the A_{CR} and B_{CR} subunits and 'tight' form, respectively. Therefore, DF complex binding seemed to change the bound form ($A_CB_{O'}$) of eA_3B_3 to the tight form ($A_{CR}B_{CR}$) by interacting with several Eh-A and -B residues (Fig. 2g, h and Supplementary Video 2).

These observations raise an intriguing question about the nature of the tight form, which occurs from the bound form by interaction with the DF complex. Figure 3a shows the nucleotide-binding site of the tight form ($A_{CR}B_{CR}$) of eV_1, which was compared with that of the bound form ($A_CB_{O'}$) of eA_3B_3. In particular, the Arg-finger (Arg 350) of B_{CR} approached 1.7 Å closer to Arg 262 relative to that of $B_{O'}$ (Fig. 3a, blue dotted line). Thus, obtaining the structure of the nucleotide-bound V_1-ATPase is essential to understanding the effect of Arg-finger movement to the nucleotide. Crystals of eV_1 were soaked in crystallization buffer with 200 μM AMP-PNP:Mg (a concentration sufficient to inhibit activity), and the crystal structure of the nucleotide-bound V_1-ATPase (denoted as bV_1) was solved to a resolution of 2.7 Å (Supplementary Table 2). The overall structure was very similar to that of eV_1 (r.m.s.d. = 0.913 Å). A strong electron density peak for AMP-PNP:Mg was observed in the binding site of the bound and tight forms, but not in the empty form (Supplementary Fig. 15), indicating that the empty form has a low affinity for AMP-PNP:Mg, consistent with the observation of the empty form in bA_3B_3.

The γ-phosphate of AMP-PNP and Mg^{2+} were fixed by the Eh-A side chains of Lys 238, Thr 239 and Arg 262 and the Arg-finger (Arg 350) of Eh-B, similar to those of bA_3B_3 (Fig. 3 and Supplementary Video 3). Superimposition of the binding site of the tight form onto that of the bound form in bV_1 revealed small but significant differences between the two sites (Fig. 3b–e). In the tight form, movement of the Arg-finger (Arg 350) 1.6 Å closer to the γ-phosphate relative to the bound form (Fig. 3d, blue dotted line) caused the rotation of the AMP-PNP γ-phosphate. Subsequently, the γ-phosphate moved 0.7 Å closer to the conserved Glu 261 (see Supplementary Video 4), which is a crucial residue for hydrolysis of yeast V_1-ATPase[29]; the corresponding residue of F_1-ATPase interacts with the γ-phosphate oxygen of ATP via a water molecule and is assumed to cleave the β-γ bond of ATP directly[6,9,10,30]. These findings suggest that hydrolysis of ATP is stimulated by this approach triggered by movement of the Arg-finger, which is induced by extensive protein–protein interactions between the DF complex and the C-terminal domains of Eh-A and Eh-B. To confirm the importance of the Arg-finger of Eh-B in ATP hydrolysis, we constructed three site-directed mutants of the Arg-finger and examined their biochemical properties (see Supplementary Table 3). The biochemical findings are consistent with the structural findings, in which hydrolysis of ATP is predicted to occur in the tight form induced by the Arg-finger approaching towards ATP. Therefore, we concluded that the obtained structure of bV_1 represents an intermediate state of waiting for ATP hydrolysis in the catalytic cycle of V_1-ATPase.

The structures and conformational differences of the three Eh-A or B subunits are apparently different from those of the F_1-ATPases, although the nucleotide-binding-site structures of these ATPases are highly conserved (see Supplementary Figs 16, 17 and Supplementary Discussion). Here we summarize a possible model of the rotation mechanism of V_1-ATPase based on the asymmetric crystal structures in this study. Figure 4a shows the structure of the C-terminal domain surface of bV_1 viewed from the top, in which two ATP:Mg molecules are bound in the bound and tight forms. Bound ATP in the tight form is awaiting ATP hydrolysis as described above. Hydrolysis of ATP seems to initiate the reaction as a trigger. New ATP molecules are unable to bind to the empty form because of its low affinity for ATP. Therefore, to continue the reaction, certain structural changes in the tight form should be induced by the conversion to ADP and Pi. If the effects of DF binding are ignored, the conformation of A_3B_3 in V_1-ATPase may change (return) to eA_3B_3 (ground structure of A_3B_3 complex) in a cooperative manner, as shown in Fig. 4b: the bound form remains stabilized by the bound ATP:Mg, the empty and tight forms in V_1 may change to the bindable form able to bind ATP and the empty form with low affinity for ATP, respectively. However, the extensive protein–protein interactions between the DF and $A_{CR}B_{CR}$ pair (tight form) may prevent this conformational change within the A_3B_3, indicating that an intermediate state should exist in place of the state of Fig. 4b. In the next step, the rotation of the DF complex seems

to be induced by ATP binding to the bindable form, or to the corresponding conformation in the intermediate state in which Eh-B, with its Arg-finger, seems to adopt an open conformation similar to B_O to enable ATP binding. Subsequently, the conformation changes to bA_3B_3, which binds ATP:Mg molecules in two bound forms, and the DF complex rotates (Fig. 4c). Finally, the older bound form changes to the tight form, induced by DF binding (Fig. 4d and Supplementary Video 3). Simultaneously, hydrolysis of ATP is again enhanced by the approach of the Arg-finger caused by the conformational change, and the enzyme reverts to its initial state, as in Fig. 4a. To understand the rotational mechanism of V_1-ATPase more precisely, further investigations should be undertaken, such as additional structural studies, molecular simulations and single-molecule observation of the rotation.

METHODS SUMMARY

Sample preparation. The A_3B_3 and DF complexes of *E. hirae* were expressed using an *Escherichia coli* cell-free protein expression system and purified as previously described[2,3]. Eh-V_1 (A_3B_3DF) was purified by gel filtration after incubation of Eh-A_3B_3 with an excess concentration (fivefold) of Eh-DF.

Crystallization, data collection and structure determination. Crystals of nucleotide-free A_3B_3 (eA_3B_3), nucleotide-bound A_3B_3 (bA_3B_3), nucleotide-free V_1 (eV_1) and nucleotide-bound V_1 (bV_1) were grown by sitting drop vapour diffusion method under the conditions described in the Methods. Diffraction data were collected from a single cryo-cooled crystal on BL41XU at SPring-8 (Harima, Japan) and NW12A, NE3A and BL1A at Photon Factory (Tsukuba, Japan). The structures of eA_3B_3, bA_3B_3 and bV_1 were solved by molecular replacement using the crystal structures of *T. thermophilus* A_3B_3 complex (PDB accession 3GQB)[19], A_3B_3 part in eV_1, and whole eV_1 as search models. The structure of eV_1 was solved by molecular replacement with single-wavelength anomalous diffraction using the structures of eA_3B_3 and Eh-DF (PDB accession 3AON), which were superimposed onto *T. thermophilus* V_1-ATPase (PDB accession 3A5C)[20]. Data collection and refinement statistics are summarized in Supplementary Tables 1 and 2.

Full Methods and any associated references are available in the online version of the paper.

Received 1 May; accepted 8 November 2012.

Published online 13 January 2013; corrected online 30 January 2013 (see full-text HTML version for details).

1. Forgac, M. Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. *Nature Rev. Mol. Cell Biol.* **8,** 917–929 (2007).
2. Arai, S. *et al.* Reconstitution *in vitro* of the catalytic portion ($NtpA_3$-B_3-D-G complex) of *Enterococcus hirae* V-type Na^+-ATPase. *Biochem. Biophys. Res. Commun.* **390,** 698–702 (2009).
3. Saijo, S. *et al.* Crystal structure of the central axis DF complex of the prokaryotic V-ATPase. *Proc. Natl Acad. Sci. USA* **108,** 19955–19960 (2011).
4. Walker, J. E. ATP synthesis by rotary catalysis (Nobel Lecture). *Angew. Chem. Int. Edn Engl.* **37,** 2308–2319 (1998).
5. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nature Rev. Microbiol.* **5,** 892–899 (2007).
6. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. *Nature* **370,** 621–628 (1994).
7. Menz, R. I., Walker, J. E. & Leslie, A. G. Structure of bovine mitochondrial F_1-ATPase with nucleotide bound to all three catalytic sites: implications for the mechanism of rotary catalysis. *Cell* **106,** 331–341 (2001).
8. Kagawa, R., Montgomery, M. G., Braig, K., Leslie, A. G. W. & Walker, J. E. The structure of bovine F_1-ATPase inhibited by ADP and beryllium fluoride. *EMBO J.* **23,** 2734–2744 (2004).
9. Bowler, M. W., Montgomery, M. G., Leslie, A. G. W. & Walker, J. E. Ground state structure of F_1-ATPase from bovine heart mitochondria at 1.9 Å resolution. *J. Biol. Chem.* **282,** 14238–14242 (2007).
10. Kabaleeswaran, V., Puri, N., Walker, J. E., Leslie, A. G. W. & Mueller, D. M. Novel features of the rotary catalytic mechanism revealed in the structure of yeast F_1 ATPase. *EMBO J.* **25,** 5433–5442 (2006).
11. Kabaleeswaran, V. *et al.* Asymmetric structure of the yeast F_1 ATPase in the absence of bound nucleotides. *J. Biol. Chem.* **284,** 10546–10551 (2009).
12. Stock, D., Leslie, A. G. & Walker, J. E. Molecular architecture of the rotary motor in ATP synthase. *Science* **286,** 1700–1705 (1999).
13. Shirakihara, Y. *et al.* The crystal structure of the nucleotide-free $\alpha3\beta3$ subcomplex of F_1-ATPase from the thermophilic *Bacillus* PS3 is a symmetric trimer. *Structure* **5,** 825–836 (1997).
14. Cingolani, G. & Duncan, T. M. Structure of the ATP synthase catalytic complex (F_1) from *Escherichia coli* in an autoinhibited conformation. *Nature Struct. Mol. Biol.* **18,** 701–707 (2011).
15. Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. Direct observation of the rotation of F_1-ATPase. *Nature* **386,** 299–302 (1997).
16. Yasuda, R., Noji, H., Yoshida, M., Kinosita, K. & Itoh, H. Resolution of distinct rotational substeps by submillisecond kinetic analysis of F_1-ATPase. *Nature* **410,** 898–904 (2001).
17. Adachi, K. *et al.* Coupling of rotation and catalysis in F_1-ATPase revealed by single-molecule imaging and manipulation. *Cell* **130,** 309–321 (2007).
18. Toei, M. *et al.* Dodecamer rotor ring defines H^+/ATP ratio for ATP synthesis of prokaryotic V-ATPase from *Thermus thermophilus. Proc. Natl Acad. Sci. USA* **104,** 20256–20261 (2007).
19. Maher, M. J. *et al.* Crystal structure of A_3B_3 complex of V-ATPase from *Thermus thermophilus. EMBO J.* **28,** 3771–3779 (2009).
20. Numoto, N., Hasegawa, Y., Takeda, K. & Miki, K. Inter-subunit interaction and quaternary rearrangement defined by the central stalk of prokaryotic V_1-ATPase. *EMBO Rep.* **10,** 1228–1234 (2009).
21. Imamura, H. *et al.* Rotation scheme of V_1-motor is different from that of F_1-motor. *Proc. Natl Acad. Sci. USA* **102,** 17929–17933 (2005).
22. Murata, T., Igarashi, K., Kakinuma, Y. & Yamato, I. Na^+ binding of V-type Na^+-ATPase in *Enterococcus hirae. J. Biol. Chem.* **275,** 13415–13419 (2000).
23. Murata, T., Yamato, I., Kakinuma, Y., Leslie, A. G. W. & Walker, J. E. Structure of the rotor of the V-Type Na^+-ATPase from *Enterococcus hirae. Science* **308,** 654–659 (2005).
24. Murata, T. *et al.* Ion binding and selectivity of the rotor ring of the Na^+-transporting V-ATPase. *Proc. Natl Acad. Sci. USA* **105,** 8607–8612 (2008).
25. Mizutani, K. *et al.* Structure of the rotor ring modified with N,N-dicyclohexylcarbodiimide of the Na^+-transporting vacuolar ATPase. *Proc. Natl Acad. Sci. USA* **108,** 13474–13479 (2011).
26. Murata, T., Yamato, I. & Kakinuma, Y. Structure and mechanism of vacuolar Na^+-translocating ATPase from *Enterococcus hirae. J. Bioenerg. Biomembr.* **37,** 411–413 (2005).
27. Yamamoto, M. *et al.* Interaction and stoichiometry of the peripheral stalk subunits NtpE and NtpF and the N-terminal hydrophilic domain of NtpI of *Enterococcus hirae* V-ATPase. *J. Biol. Chem.* **283,** 19422–19431 (2008).
28. Zhou, M. *et al.* Mass spectrometry of intact V-type ATPases reveals bound lipids and the effects of nucleotide binding. *Science* **334,** 380–385 (2011).
29. Liu, Q. *et al.* Site-directed mutagenesis of the yeast V-ATPase A subunit. *J. Biol. Chem.* **272,** 11750–11756 (1997).
30. Dittrich, M., Hayashi, S. & Schulten, K. On the mechanism of ATP hydrolysis in F_1-ATPase. *Biophys. J.* **85,** 2253–2266 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. E. Walker for his suggestions, especially through the structural studies of F_1-ATPase. The synchrotron radiation experiments were performed at SPring-8 and Photon Factory (proposals 2008S2-001, 2011S2-005, 2009G660, 2009B1031 and 2012G132). We also thank the beamline staff at BL41XU of SPring-8 (Harima, Japan) and NE3A, NW12A and BL1A of Photon Factory (Tsukuba, Japan) for help during data collection. This work was supported by the Targeted Proteins Research Program, grants-in-aid (23370047, 23118705), Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government.

Author Contributions T.M. designed the study. S.A., Y.K. and N.O. constructed DNAs. Y.I.-K., T.T. and M.S. expressed and purified the proteins. K.S. and T.M. crystallized the proteins. S.A., S.S., K.S., K.M. and T.M. collected X-ray data. S.A., S.S. and K.M. processed and refined X-ray data. S.A. and K.S. performed functional analysis. S.A., S.S., I.Y. and T.M. analysed the results. S.A. and S.S. prepared figures and videos. T.M. wrote the paper. All authors discussed the results and commented on the manuscript. The study was managed by S.Y., S.I., I.Y. and T.M.

Author Information Atomic coordinates and structure factors for the A_3B_3 and V_1-ATPase complexes have been deposited in the Protein Data Bank under the accession codes 3VR2 (nucleotide-free A_3B_3 at 2.8 Å), 3VR3 (nucleotide-bound A_3B_3 at 3.4 Å), 3VR4 (nucleotide-free V_1-ATPase at 2.2 Å), 3VR5 (nucleotide-free V_1-ATPase at 3.9 Å) and 3VR6 (nucleotide-bound V_1-ATPase at 2.7 Å). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.M. (t.murata@faculty.chiba-u.jp).

METHODS

Protein preparation. An *Escherichia coli* cell-free protein expression system, as described elsewhere[31], was used to synthesize the Eh-A_3B_3 and DF complexes using a mixture of plasmids containing the corresponding genes. The expressed complexes were purified as previously described[2,3]. Selenomethionine-labelled Eh-A_3B_3 and DF complexes were also prepared to facilitate X-ray structure determination. The sum of ATP and ADP contamination in 0.5 μM purified Eh-A_3B_3 and DF complexes (denatured with 0.6 M perchloric acid) was estimated with a luciferin–luciferase assay[32], after conversion of ADP to ATP using an ATP regeneration system[3] for 30 min at room temperature; ATP and ADP contamination was undetectable (less than 0.1 nM). The V_1-ATPase (Eh-A_3B_3DF) was reconstituted and purified as follows: purified Eh-A_3B_3 and Eh-DF in buffer A (20 mM Tris-HCl, pH 8.0, 150 mM NaCl and 2 mM dithiothreitol (DTT)) were mixed at a 1:5 molar ratio with the addition of MES (100 mM final concentration; pH 6.0), and were incubated with and without 0.2 mM AMP-PNP and 5 mM $MgSO_4$ for 1 h. Reconstituted V_1-ATPase with and without AMP-PNP:Mg was purified using a HiLoad 16/60 Superdex 200 (GE Healthcare) column equilibrated with buffer B (20 mM MES, pH 6.5, 10% glycerol, 100 mM NaCl, 5 mM $MgSO_4$ and 2 mM DTT), respectively. Purified complexes were concentrated with an Amicon Ultra 30 K unit (Merck Millipore).

Protein crystallization. All crystallization trials were performed using the sitting-drop vapour diffusion method at 296 K. The crystals were soaked in cryoprotectant by incrementally increasing the glycerol concentration to 20%. The crystals were then mounted on cryo-loops (Hampton Research), flash-cooled and stored in liquid nitrogen.

(1) Eh-A_3B_3 without nucleotide (eA_3B_3): Eh-A_3B_3 crystals were obtained by mixing 0.5 μl protein solution (12 mg ml^{-1} protein in buffer A) with 0.5 μl reservoir solution (0.1 M MES-Tris, pH 8.5, 24% PEG-3350 and 0.2 M ammonium acetate).

(2) Eh-A_3B_3 with AMP-PNP:Mg (bA_3B_3): Eh-A_3B_3 crystals were obtained by mixing 0.5 μl protein solution (11 mg ml^{-1} protein in buffer A) supplemented with 5 mM AMP-PNP and 5 mM $MgSO_4$ with 0.5 μl reservoir solution (0.1 M HEPES, pH 7.5, 26% PEG-3350, and 0.2 M sodium chloride).

(3) V_1-ATPase with AMP-PNP:Mg (eV_1): V_1 crystals were obtained by mixing 0.5 μl protein solution (12 mg ml^{-1} protein in the presence of 0.2 mM AMP-PNP in buffer B) with 0.5 μl reservoir solution (0.1 M Bis-Tris propane, pH 6.5, 19% PEG-3350 and 0.2 M sodium fluoride).

(4) V_1-ATPase without nucleotide (eV_1(L)): V_1 crystals were obtained by mixing 0.5 μl protein solution (10 mg ml^{-1} protein in buffer B) with 0.5 μl reservoir solution (0.1 M Bis-Tris propane, pH 6.5, 20% PEG-3350 and 0.2 M sodium fluoride).

(5) V_1-ATPase soaked with AMP-PNP:Mg (bV_1): V_1 crystals that were obtained in (3) were soaked for 5 h in 0.1 M Bis-Tris propane, pH 6.5, 21% PEG-3350, 0.2 mM AMP-PNP (a concentration sufficient to inhibit the activity), 3 mM $MgSO_4$, 0.2 M sodium chloride (replaced for sodium fluoride) and 20% glycerol.

Structure determination. All X-ray diffraction data were collected from a single crystal at a cryogenic temperature (100 K).

(1) Eh-A_3B_3 without nucleotide (eA_3B_3): X-ray diffraction data were collected on beamline BL41XU ($\lambda = 1.0000$ Å) at SPring-8 (Harima, Japan). The collected data were processed to 2.8 Å using iMosflm[33] and then scaled by Scala from the *CCP4* program suite[34]. The structure was solved by molecular replacement with MOLREP[35] using the poly-Ser model of A_3B_3 complex from *T. thermophilus* (PDB accession 3GQB)[19] as a search model.

(2) Eh-A_3B_3 with AMP-PNP:Mg (bA_3B_3): X-ray diffraction data were collected on beamline BL41XU ($\lambda = 1.0000$ Å) at SPring-8. The collected data were processed to 3.4 Å using iMosflm[33] and then scaled by Scala[34]. The structure was solved by molecular replacement with MOLREP[35] using the structure of Eh-A_3B_3 in eV_1 as a search model.

(3) V_1-ATPase with AMP-PNP:Mg (eV_1): X-ray diffraction data were collected on beamline NW12A ($\lambda = 0.97919$ and 1.0000 Å) at Photon Factory (Tsukuba, Japan). The collected data were processed and scaled to 2.6 and 2.2 Å using XDS[36]. The structure was solved by MR–SAD (molecular replacement with single-wavelength anomalous diffraction) using Phaser[37]. The partially refined A and B subunits from the structure of Eh-A_3B_3 without nucleotide and the DF complex from *E. hirae* (PDB accession 3AON) were superimposed onto the V_1-ATPase from *T. thermophilus* (PDB accession 3A5C)[20]. The superimposed model was used as an initial search model. The overall figure of merit (FOM $= \Sigma P(\alpha)\exp(i\alpha)/\Sigma P(\alpha)$, in which $P(\alpha)$ is the probability distribution for the phase(α)) was 0.45 using combined phases of SAD from selenium and molecular replacement at 3.0 Å.

(4) V_1-ATPase without nucleotide (eV_1(L)): X-ray diffraction data were collected on beamline NE3A ($\lambda = 1.0000$ Å) at Photon Factory. The collected data were processed to 3.9 Å using XDS. The structure was solved by molecular replacement with Phaser using the crystal structure of eV_1 as a search model.

(5) V_1-ATPase soaked with AMP-PNP:Mg (bV_1): X-ray diffraction data were collected on beamline BL1A ($\lambda = 1.0000$ Å) at Photon Factory. The collected data were processed to 2.7 Å using HKL2000 software (HKL Research). The structure was solved by molecular replacement with MOLREP using the crystal structure of eV_1 as a search model.

The atomic models were manually built using Coot[38] and iteratively refined using REFMAC5[39] (REFMAC5 and PHENIX[40] were used for refinement of eV_1). The refined structures were validated with PROCHECK[41] and RAMPAGE[42]. The crystallographic and refinement statistics are summarized in Supplementary Tables 1 and 2. The r.m.s.d. values of superimpositions for each Eh-A, Eh-B or Eh-AB pair in the crystal structures of A_3B_3 and V_1-ATPase are listed in Supplementary Tables 4 and 5. Figures were prepared using PyMOL (The PyMOL Molecular Graphics System, Version 1.3, Schrodinger, LLC.)

Characterization of the Arg-finger mutants. Mutagenesis of the Arg-finger (Arg350Ala/Glu/Lys) of Eh-B was performed using the QuikChange site-directed mutagenesis kit (Agilent Technologies). The Arg350Ala Eh-B mutant could not be purified as a complex with Eh-A. The other mutants were purified as described above. Binding affinities between the Eh-A_3B_3 mutants and the DF complex were measured using the Biacore T100 system (GE Healthcare) as described previously[3]. The dissociation constant (K_d) was determined using BIAevaluation software (version 1.1), which uses the Langmuir isotherm model that assumes a 1:1 binding stoichiometry. ATPase activity was measured according to a previous report using the ATP-regeneration system[3].

31. Kigawa, T. *et al.* Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. *J. Struct. Funct. Genomics* **5,** 63–68 (2004).
32. Deluca, M. & McElroy, W. D. Purification and properties of firefly luciferase. *Methods Enzymol.* **57,** 3–15 (1978).
33. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67,** 271–281 (2011).
34. Collaborative Computational Project, Number 4. The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
35. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66,** 22–25 (2010).
36. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
37. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
39. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
40. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
41. Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **8,** 631–639 (1998).
42. Lovell, S. C. *et al.* Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins* **50,** 437–450 (2003).

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature11736

Corrigendum: Resolving the evolutionary relationships of molluscs with phylogenomic tools

Stephen A. Smith, Nerida G. Wilson, Freya E. Goetz, Caitlin Feehery, Sónia C. S. Andrade, Greg W. Rouse, Gonzalo Giribet & Casey W. Dunn

Nature **480,** 364–367 (2011); doi:10.1038/nature10526

In this Letter, we investigated the evolutionary relationships of molluscs with multigene data sets assembled from new transcriptome data and published genomes and transcriptomes. Since publishing these results, examination of our gene sequence matrix by others revealed that all instances of six amino acids (E, F, I, L, P and Q) were replaced by ambiguous characters in our super matrix. This led to the exclusion of data that should have been in the final analyses. The data exclusion was caused by incorrect handling of protein data at the final stage of matrix concatenation by the published program Phyutility (http://code.google.com/p/phyutility/). We have fixed the program, regenerated the final matrices, and re-run our analyses. There was minimal impact on our results, with no changes in the topology of the tree at deep nodes that had consistent strong support in our published analyses. There are minor variations in support values in the corrected analyses. The corrected matrices have been deposited at Dryad under the existing accession number (http://dx.doi.org/10.5061/dryad.24cb8). Figure 1 shows the corrected Fig. 2 of the original Letter, with corrected support values. In the text of the original Letter, the sentence "Bayesian analyses using the site-heterogeneous CAT model of protein evolution also place Scaphopoda as the sister group to Gastropoda, with a posterior probability of 89%" should read "Bayesian analyses using the site-heterogeneous CAT model of protein evolution also place Scaphopoda as the sister group to Bivalvia, with a posterior probability of 81%". In the Methods Summary, both instances of "27%" in the following phrase should be "41%": "27% character occupancy (that is, 27% of the matrix consists of unambiguous amino acid data, with the remainder being missing data or alignment gaps)" and "21%" should be "32%" in the sentence "This matrix has 40% gene occupancy, 21% character occupancy and is 216,402 sites long.". In the Methods, the following two sentences should be removed: "PhyloBayes misidentified the data type of our matrix as DNA, resulting in model misspecification and lack of convergence. We conducted the analyses presented here with a modified version that was forced to read all matrices as protein sequences." The next three sentences from the final paragraph of the Methods should now read "Five PhyloBayes runs under the fully parameterized CAT model were run, and each converged by 2,000 cycles based on time series plots of the likelihood scores and number of partitions. The runs were allowed to run, each for more than 3,500 cycles, and estimated about 300 categories for the model." instead of "Five PhyloBayes runs under the fully parameterized CAT model each converged at around 1,500 cycles (at least 86,000 generations) based on time-series plots of the likelihood scores and number of partitions. The runs were allowed to run for 5,000 cycles for two runs and 2,500 cycles for three runs. The runs estimated 140 (±10) categories for the model." The original Letter's Supplementary Figures 2–9 have also been updated with the results of analyses based on the corrected matrices. Differences in results for these figures include increased support for some relationships highlighted in the original manuscript, and changes in some relationships within Bivalvia. These changes do not alter any of the conclusions of our manuscript. We thank Hervé Philippe, Raphael Poujol and Béatrice Roure for bringing this error to our attention.

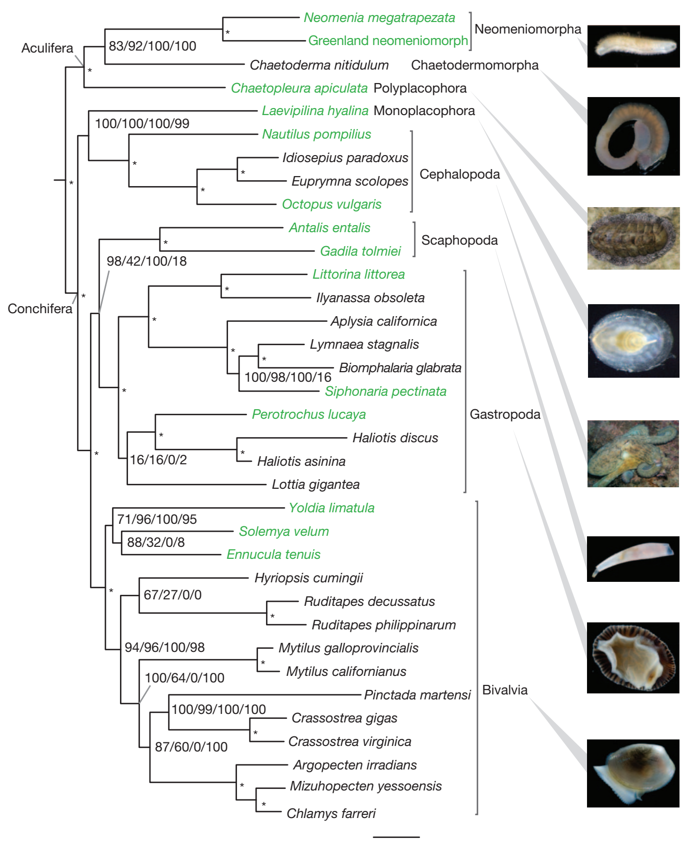


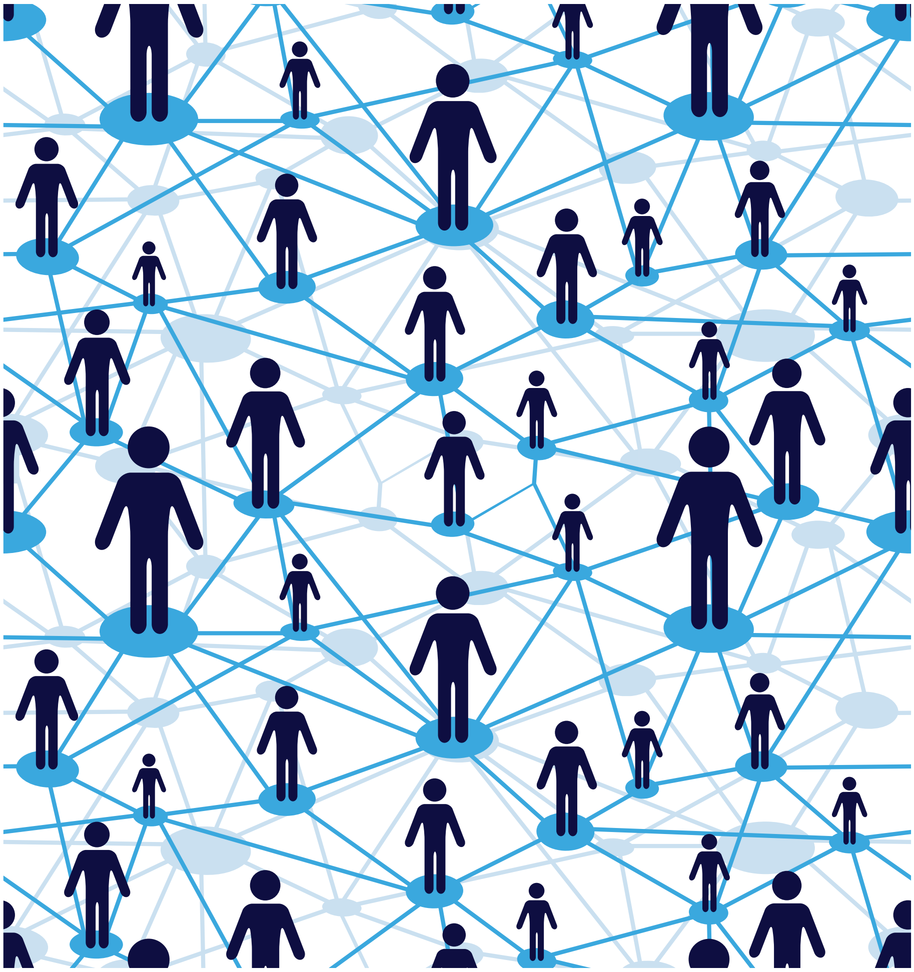
Figure 1 | This is the corrected Fig. 2 of the original Letter.

CAREERS

TURNING POINT Ecologist's open notebook leads to opportunities **p.711**

MISCONDUCT Male scientists commit fraud more often than women, says study **p.711**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**



SWEET LANA/SHUTTERSTOCK

COLUMN

Better connected

Informal networks are key to idea-sharing, argue **Mark Fishman**, **Robert Cross** and **Brigitta Tadmor**.

Research-and-development companies are constantly changing their organizational structures to nurture innovation and increase productivity. Yet no single organizational model has emerged as the best option in either academia or industry.

Formal organizational structures are important, but scientific discovery is an emotional process shaped by social environment. In 2011, to improve understanding of interpersonal dynamics and to foster innovation, the pharmaceutical company Novartis, based in Basel, Switzerland, conducted a study of employee interactions in its global drug-discovery arm: the Novartis Institutes for Biomedical Research (NIBR). At the time of the study, the NIBR employed about 6,600 people at ten sites in the United States, Europe and Asia.

We wanted to understand how influential individuals ('key nodes') in the network affect information access, idea-sharing, problem-solving and other aspects of scientific discovery. We found that informal network structure — how people interact on a daily basis — differs dramatically from formal, hierarchical organizational structure. Key nodes in the informal network are better at sharing ideas and resolving difficulties than are formally appointed managers or leaders. Our data also suggest that improving how a person interacts with his or her informal network might have a greater impact on innovation than mandatory leadership courses and performance review. These findings may apply to many research-and-development organizations.

NETWORK ARCHITECTURE

To assess the NIBR's network of researchers, we sent out a questionnaire asking respondents to identify others in the organization with whom they needed to interact, whether in person, on the phone or by e-mail. Every interaction identified had to be crucial to the respondent's work. We asked respondents to assess the reasons for, and quality of, these interactions, and whether they needed more access to certain people, departments or areas of expertise. We then categorized the reasons for interactions: information access, problem-solving, idea-sharing, access to leaders and decision-makers, political support and personal support or advice. We categorized interactions as energizing or de-energizing; facilitating or hindering the discussion of new ideas and divergent points of view; providing a sense of purpose or urgency; and modelling leadership behaviours.

Finally, we collected demographic data on each person, including the length of their tenure at the NIBR, and their main language, gender and personality type according to a Myers–Briggs assessment. The overall participation rate was almost 70%, with about 70,000 reported connections (an average of 11 per employee).

Around 60% of all described interactions occurred within a department; the remaining 40% spanned scientific disciplines. Many scientists reported a need for greater access to other researchers or leaders, both within and across departments. It seems that network size depends on rank and tenure: the networks of the highest-ranking leaders (0.5% of the organization) tend to be almost ten times larger than those of people in entry-level positions, and it takes about three years from entering the ▶

organization for a person's network to reach average size. Researchers — particularly in Asia — interact less frequently with colleagues in different geographical areas or even in different buildings or floors of the same building. Other research organizations have reported[1] that interactions drop exponentially beyond distances greater than 15 metres.

Finally, we observed subtle network preferences based on culture, language and gender. For example, scientists in Shanghai, China, have smaller networks than those in the United States or Europe but spend twice as much time in each interaction. This supports the idea that relationship-building is an important part of the culture of Asian firms. We also observed that scientists who speak the dominant language at a site have larger networks than scientists who use the minority language; furthermore, men have larger networks than women, preferentially interact with other men and tend not to consider women as leadership role models. These findings are consistent with the sociological concept that people favour members of their own 'group' over others.

KEY NODES

Overall, we found that scientists who have positive interactions with others have larger networks than would be predicted from their formal position in the hierarchy (and vice versa). Those who can instil a sense of purpose and inspire others are pursued by their network for idea-sharing, information access, problem solving and personal support. We teased out three distinct categories of scientists who act as key nodes in their networks:

Experts These scientists offer expertise in certain technical, scientific or clinical areas across the NIBR. There are experts at all levels and across all functions in the organization; some offer their expertise locally (within a subunit of a department) whereas others are more widely connected, providing expertise across geographies and disciplines. As individuals, they tend to be analytical and introverted.

Mentors These scientists provide others with a sense of purpose, and their colleagues feel comfortable approaching them with new ideas and divergent points of view. As a result, mentors are sought for help with problem-solving and for personal support and advice (see 'Network nodes'). They exist at all levels in the organization but their positive interactions make their networks 50% larger than the average among their peers. No single personality type is dominant among these individuals.

Brokers These scientists have large networks and are connected broadly across functions and geographies. They tend to be high ranking and visible in the organization, and they mainly provide political support and access to decision-makers. They are not sought primarily for idea-sharing, problem-solving or scientific expertise. They tend to have extroverted and assertive personalities.

The effect of key nodes on the organization is powerful. As reported in previous studies, one important distinction is whether a person energizes or de-energizes people in his or her network[2]. We found that people who can get colleagues motivated also energize their networks. Furthermore, energizers create an environment that fosters collaborations and encourages joint problem-solving and idea-sharing. De-energizers, by contrast, create an environment in which people are reluctant to collaborate and share ideas, and in which interactions are perceived as demotivating. Mentor nodes are invariably energizing, whereas expert and broker nodes can have either effect. Both positive and negative effects are more pronounced when people have large networks.

THE IMPLICATIONS

How did we make use of the data that we gathered? The NIBR study was done anonymously to enable a high rate of response, which hindered any direct, open intervention with specific individuals. But we did take some action. First, we shared the study's general themes and observations with the organization, highlighting the level and quality of interactions in each department and between departments, and identifying areas that lacked intra- or interdepartmental collaboration. This gave leaders an opportunity to address these issues.

Second, using a confidential website, we provided each person with information about how they were perceived by the network. We offered individual coaching and workshops for small groups or teams, including workshops on personality styles and subconscious biases, and how these factors affect interactions with others.

"Those who can instil a sense of purpose and inspire others are pursued by their network for idea-sharing and problem-solving."

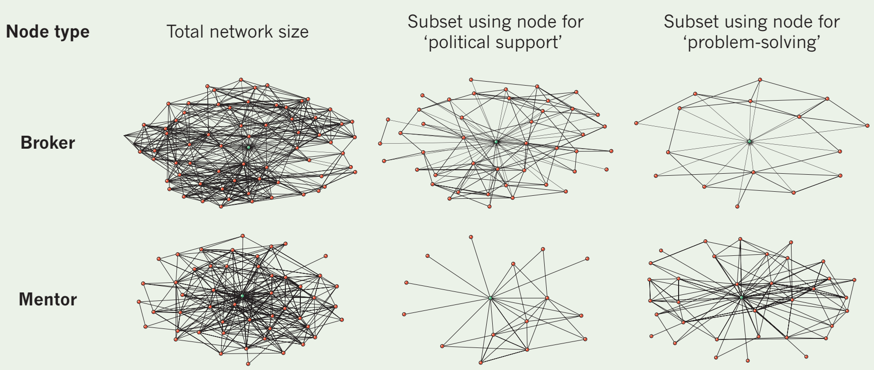
More than 60% of all people at the NIBR accessed their personalized, web-based network information, including people who didn't respond to the survey but were named in other people's networks. More than 10% (about 700 people) voluntarily engaged in follow-on activities (among high-ranking leaders, that figure was 25%), including those with smaller than expected or poor-quality networks.

We believe that helping a relatively small group of self-motivated scientists to improve their interactions — by becoming easier to approach with new ideas, for example — will create an innovative culture much more effectively than making formal changes in the organizational structure or mandating training for managers or leaders. And other network research suggests the same[3].

Information networks are often not considered in traditional performance evaluation at science organizations, in which line managers make assessments with little input from others. We believe that such omissions fail to provide incentives for open collaboration. Our mapping of interactions at the NIBR suggests that, whatever the field of science, feedback from the informal network, coupled with individual and small-group coaching, will facilitate a creative, innovative culture. ■

NETWORK NODES

Brokers and mentors (green) both work closely with lots of people (red), but whereas brokers are mainly sought out for political support, mentors are in demand for problem-solving.



SOURCE: NIBR

Mark Fishman *is president of the Novartis Institutes for Biomedical Research (NIBR) in Cambridge, Massachusetts.* **Robert Cross** *is an associate professor of management at the University of Virginia in Charlottesville.* **Brigitta Tadmor** *is vice-president of education, diversity and inclusion at the NIBR.*

1. Allen, T. J. *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R&D Organisation* (MIT Press, 1977).
2. Baker, W., Cross, R. & Wooten, M. in *Positive Organizational Scholarship: Foundations of a New Discipline* (eds Cameron, K. S., Dutton, J. E. & Quinn, R. E.) 328–342 (Berrett-Koehler, 2003).
3. Xie, J. *et al. Phys. Rev. E* **84,** 011130 (2011).

TURNING POINT
Carl Boettiger

While working towards his PhD in mathematical ecology at the University of California, Davis, Carl Boettiger launched a kind of career experiment. He started keeping an open lab notebook online, aiming to find new ways to communicate science and to seed collaborations. As he begins a postdoc at the University of California, Santa Cruz, Boettiger explains the pros and cons of sharing his work on the Internet.

Your first degree was in physics. How did you move to ecology?
As an undergraduate, I didn't enjoy biology classes — basically memorizing facts — as much as physics classes, which involved problem-solving. But I knew I that I would leave physics after I got a quantitative background. In my second year, I met Simon Levin, a mathematical ecologist at Princeton University in New Jersey. I believe I told him that I wanted to go into ecology to make it more mathematical. He, I'm sure, resisted the urge to throw me out of the window. But he invited me to join his weekly lab presentations. That and a research project with one of his postdocs were my only ecology training before my PhD.

Have you had a turning point in your career?
I got a computational-science fellowship from the US Department of Energy, designed to encourage supercomputing throughout the sciences. Getting that and thinking in a more computational way were big turning points. The programme is very interdisciplinary. It brings us — astrophysicists, genome biologists, ecologists — to a conference every year so that we can talk in the same computational language, if not the same scientific language.

Why did you start your open notebook?
I didn't have much research training. While trying to figure out what I had done months before, I realized that I should be more organized. I stumbled on tips for keeping an electronic notebook, and started mine in January 2010. I wanted to see if it could help me to educate people about my work, communicate faster with colleagues or make my research more transparent and reproducible.

How many people check out your work?
I get about 60,000 page views each year — roughly 150 a day, usually from 50 visitors.

What has keeping the notebook taught you?
I discovered an online community, including ecologists, that I didn't know existed. I have had tonnes of fruitful interactions. I have been given valuable feedback, including suggestions on tackling computational problems, and I participate in conversations about how to make large-scale modelling more reproducible. I was surprised to see colleagues emulate my approach, and almost terrified when other graduate students whom I had never met started keeping open notebooks. I thought, "This is still an experiment; I hope I'm not responsible for anything" — whether it was harsh criticism or someone getting scooped.

How have advisers and colleagues reacted?
Usually more with ambivalence than discouragement. One of the biggest challenges is addressing data-sharing concerns. We have to establish whether collaborators are comfortable putting their work in this environment. If someone is uncomfortable, but hasn't said explicitly not to share the work, I have to decide what to do.

Were you afraid of getting scooped?
That concern is there, but I haven't experienced any scoops. I think it is an overrated fear, especially when compared with the risk of being unknown in your field. My notebook has helped me to extend my reach in ecology and computing. People were aware of me before I had published — which led, for example, to invitations to review papers. If anything, I was reluctant to put things up in case they contained mistakes. Every mistake that I made during my PhD is in there. But if I am trying to resolve an error, I can easily show others all the work I have done and the steps I have taken, and ask them for advice. ■

INTERVIEW BY VIRGINIA GEWIN

MISCONDUCT

Fraud by gender

Men have committed research fraud more often than women, according to a study published on 22 January (F. C. Fang *et al. MBio* **4,** e00640-12). The authors reviewed 215 cases of fraud in the life sciences reported by the US Office of Research Integrity between 1994 and 2012. Of those, 65% were committed by male scientists. In cases involving US faculty members, 70% of whom are men, male researchers were responsible 88% of the time. In those involving postdocs, of whom 61% are male, it was 69%. Arturo Casadevall, a microbiologist at the Albert Einstein College of Medicine in New York, who led the study, says that the finding underscores the need for ongoing ethics training.

PHYSICS

Permanent jobs scarce

The recession seems to have increased the proportion of US physics PhD holders who are taking postdoc jobs rather than waiting for permanent positions, says a brief released on 17 January by the American Institute of Physics (AIP) in College Park, Maryland. *Physics PhDs: One Year Later*, which is based on survey data, reports that 59% of people who earned physics PhDs in the United States in 2009 and 2010 and remained in the country took postdocs within a year. Of those, 13% did so because they could not find permanent work, compared with 7% from the classes of 2007 and 2008, according to AIP data. Similarly, 44% of physics PhD holders who took temporary positions had failed to find a suitable permanent job, up from 42% from the classes of 2007 and 2008.

GENDER BIAS

Resources denied

Women need to justify requests for pay increases and other resources more than men, a study reports (H. R. Bowles and L. Babcock *Psychol. Women Quart.* http://doi.org/j99; 2013). The authors asked more than 500 university graduates with some work experience to evaluate videos of men or women asking supervisors for resources. Women's requests generally met with disapproval. Co-author Hannah Bowles, who studies gender and leadership at Harvard University in Cambridge, Massachusetts, says that mentioning a mentor or adviser in negotiations may help women. "Use a lot of 'we' language and signal that you have these positive working relationships," she says.

TO MY FATHER

Postcard from the edge.

BY DAVID G. BLAKE

JACEY

Interstellar uplink successful: 20-minute propagation delay.

This is farewell.

From your office window, I can see the colony's artificial biosphere disintegrating, fiery fragments crumbling free and bursting into showers of gold sparks. Across the broken horizon, prismatic tendrils of gas and dust bleed through the cracks, producing an array of writhing colours that span the optical spectrum. The result is remarkable.

The expanding cloud of spores, which reeks of mildew and decay, is not as impressive as the deluge of gold sparks, nor as striking as the rainbow weaves, but it *is* as exceptional in its own destructive way. It also shrouds the bodies that litter the streets below, although the memories of their faces warped with agony cannot be interred.

Unimpeded, plasmoids will spread those foul-smelling spores throughout the heliosphere. I recommend an immediate system-wide purge, followed by comprehensive tests to confirm the eradication of the radiotrophic fungi. It will do nothing for the colony, and even less for those of us left behind, but it should prevent such a disaster from recurring.

I am … *relieved* that you made it out before it was too late.

The bookshelf behind your desk still holds many of your favourite books: a few flawlessly positioned, as if nothing had changed; some crooked or upturned; others spilled out over the cold floor. You emptied the locked desk drawer — and the wall safe behind the painting of a sunset on Mars — but left the others filled with things not deemed significant enough to take. You even left behind the bottle of scotch that you were saving for a special occasion.

Shattered on the floor beside your overturned chair, an empty picture frame taunts me. I can recall every detail of the missing picture: you and Claire leaning against the model of Earth mounted outside the laboratory, little Daniel asleep in your arms; the flush of first light captured rising behind you, its erratic glow glinting along the curve of the artificial biosphere like a smear of oil on glass.

You never noticed my hard metal face — so *different* from little Daniel's — pressed against one of the upper laboratory windows; when it came to me, you failed to notice many things. You seemed so satisfied, so at peace … so whole. I could not look away. Even now, I am forced to rip my thoughts out of the grasp of that poignant memory.

From the moment you gave me life, you taught me to learn and adapt through observation and research. I embraced the process with vigour, each fresh crumb of gleaned information filling me with the pleasure of your approval. In spite of my eagerness, it required extensive research to learn what it was that I felt as I stared down at you and your new family: diminished, as though I had become nothing more than an outmoded contrivance.

Have you ever felt diminished, Father? A knot — a malignant tumour — forms in your very core. As it grows larger and larger, you become smaller and smaller. It is a harrowing feeling, a feeling that *endures*, and it carries with it the certainty that there is no limit to how insignificant you can become. I gained no pleasure from discovering such a wounded part of me.

When first I woke to find you gone, I made myself believe that there simply had not been enough time for you to take me with you.

Yet you found the time to empty the wall safe and the locked desk drawer.

You found the time to take several of your favourite books.

You found the time to take Claire.

You found the time to take little Daniel.

You even found the time to take that picture out of its shattered frame.

The world is such a fearsome, lonely place, when one is so small. How am I supposed to adapt to that?

Anger is something I learned about by observing yo—

Interstellar uplink terminated.

Remote relay module activated.

Interstellar uplink reestablished.

The rising spores forced me out of your office and onto the roof of the laboratory. I do not have much time left.

No point in wasting any of it asking questions that you will never have the opportunity to answer — not that I believe you would answer them, if you were offered such a chance. In addition, I will no longer waste time on anger, even though it feels as if gears are grinding hard against circuits inside me.

The artificial biosphere is all but gone, leaving behind a sky framed by its smouldering skeleton. Our — *my* — home is barely recognizable now. I take comfort in the knowledge that there is no one left alive to suffer through the end … no one but me. I could block the pain if I wanted to, but it makes me feel less diminished, as though pain is reserved only for those who are significant enough to have earned it.

This is farewell.

Interstellar uplink terminated. ■

David G. Blake *lives in Pennsylvania with his girlfriend and their chocolate labrador. His work has appeared in* Beneath Ceaseless Skies, Daily Science Fiction *and other publications.*

Does consumption rate scale superlinearly?

ARISING FROM S. Pawar, A. I. Dell & V. M. Savage *Nature* **486,** 485–489 (2012)

A recent paper by Pawar and colleagues[1] has provided important insights into the consequences of foraging behaviour for food-web dynamics. One notable pattern predicted by their analysis is that consumption rate (c) scales superlinearly ($c \propto m^{1.16}$) with consumer body mass (m) in three-dimensional (3D), but not two-dimensional (2D), foraging spaces. Although we feel that the authors should be applauded for this interesting contribution, we argue that their result is not consistent with established life-history theory. To resolve this contradiction, progress in both fields is probably required, including new empirical studies in which consumption rate, metabolism and dimensionality are examined directly under natural conditions.

One inconsistency is that, under a superlinear scaling of consumption rate, most models of ontogenetic growth and life-history optimization would predict infinitely large body sizes[2–6]. To obtain realistic ranges in maturation and maximum sizes, an equally superlinear scaling of metabolism and/or a positive scaling of mortality rate with body size would be required, but these are not generally found in nature[7,8]. Although biomechanical factors may set an ultimate limit to body size, such 'universal' constraints cannot account for body-size variation among species with similar body plans living in environments with similar physical properties (for example, pelagic fish). However, biomechanical constraints can explain differences in the sizes of organisms between habitats (that is, pelagic organisms can be larger than terrestrial organisms because they are less constrained by gravity) and can also explain why terrestrial organisms foraging in 3D (for example, flying or canopy-dwelling species) are more limited in size than ground-dwelling organisms that forage in 2D — a pattern that contradicts the predictions of Pawar *et al.* concerning size and dimensionality.

Another inconsistency is that substantial difference in scaling between realized and maximum consumption ($m^{0.75}$)[1,9,10] implies that the superlinear model must be violated; first, for body sizes larger than the point at which these two relationships intersect (that is, realized consumption cannot be greater than maximum consumption); and second, for body sizes smaller than the point at which growth is prohibited because realized consumption is equal to or lower than maintenance consumption (that is, the minimum consumption required to cover metabolic costs) (Fig. 1, Methods). One possible explanation for this unrealistic behaviour at small body sizes is that the model was designed for consumption rates per trophic link[1], such that in nature small consumers would meet their needs by eating additional resources. However, narrowing the focus of the model in this way would necessarily limit its ability to describe the behaviour of natural ecosystems, as most consumers eat more than one prey species and there is no reason to expect that small consumers would be more generalist than large consumers[11].

These inconsistencies raise questions regarding both model mechanisms and analyses of empirical data. First, the model does not include the ability of prey to evade predators as a function of the prey's reaction distance. Instead, capture success per predator–prey encounter is assumed as a constant[1], although it can vary by orders of magnitude in nature[12] and substantially reduce consumption rates for larger predators. Second, the consumption-rate data consist mainly of laboratory experiments, which can overestimate field consumption rates through container effects (in the 3D data, foraging arenas are proportionally smaller for larger consumers; that is, container size scales sublinearly with consumer size, see Methods) that allow large consumers to feed unhindered by fear of predation[13] while reducing the efficacy of predator evasion by prey[12]. Finally, the authors implicitly emphasize interspecific patterns, but their model is based on principles equally applicable to intra- and interspecific size variation, and both were included in their empirical data. Combining intra- and interspecific variation in one analysis (for example, in the 3D data, life stage changed systematically with species size) can introduce a bias to estimates of scaling exponents (Fig. 2, Methods).

Limitations notwithstanding, 3D consumption rates remain higher than 2D rates. Other studies have shown that metabolism scales more steeply in pelagic (3D) than in surface-dwelling (2D) animals[8,14], thus raising the fundamental question of whether consumption rates are driven internally by consumer energetic demand or externally by resource availability. In addition, the important role of predator–prey size relations in this work suggests that the effects of dimensionality

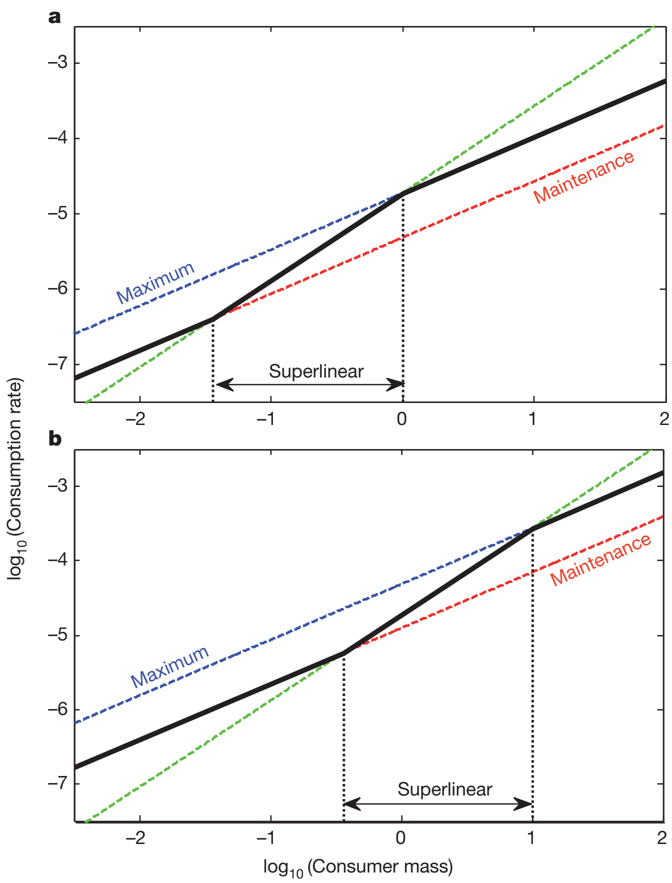


Figure 1 | Scaling of consumption rates with consumer mass. a, Superlinear scaling (green line, $c \propto m^{1.16}$)[1] is only feasible (black line) between the interceptions with maintenance and maximum consumption ($c \propto m^{0.75}$)[9]. The upper intercept was arbitrarily set at 1 kg, the lower intercept (36 g) is based on maintenance consumption of ectothermic vertebrates (see Methods). Consumption rates are in kg s^{-1}, consumer mass in kg. **b**, As maintenance consumption is a constant proportion of maximum consumption, changing the upper intercept (10 kg) has no effect on the feasible range of superlinear scaling, which encompasses only 1.4 orders of magnitude of body size (or 3.1 and 0.5 orders of magnitude in invertebrates and endothermic vertebrates, respectively).

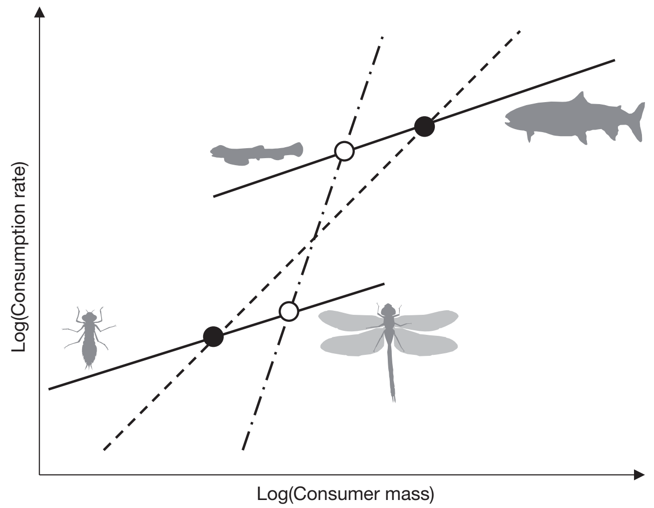


Figure 2 | Effect of biased sampling of stages of development on the scaling of consumption. Continuous lines represent intraspecific scaling, black circles mark the average or typical body mass for each species, and the dashed line is the resulting true interspecific scaling. When sampling from larger species is biased towards immature stages and vice versa, the resulting scaling estimate (dashed-dotted line) exceeds both the true intra- and interspecific scalings.

on these relations for intra- versus interspecific size variation would be a fruitful focus for future study.

METHODS SUMMARY

Maintenance consumption was estimated by dividing maximum consumption by the ecological scope (defined as maximum consumption/metabolism, whose values are typically 19.4, 3.9 and 1.6 for invertebrates, ectothermic vertebrates and endothermic vertebrates, respectively[15]). We assumed scaling exponents of 1.16 (ref. 1) and 0.75 (refs 1, 9, 10) for realized and maximum consumption.

Sublinear scaling of container size was tested by a log-log regression ($b = 0.50$, 95% CI = 0.41–0.60) on 3D data from the paper by Pawar *et al.*[1] We tested the association between species size (using invertebrates versus vertebrates as a surrogate) and life stage (juvenile versus adult) through chi-squared tests ($P = 0.27$ for 2D and $P < 0.001$ for 3D).

Henrique C. Giacomini[1], Brian J. Shuter[1,2], Derrick T. de Kerckhove[1] & Peter A. Abrams[1]

[1]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Harbord St., Toronto, Ontario M5S 3G5, Canada.
e-mail: hgiacomini@gmail.com
[2]Harkness Laboratory of Fisheries Research, Ontario Ministry of Natural Resources, 2140 East Bank Drive, Peterborough, Ontario K9J 7B8, Canada.

Received 13 August; accepted 22 November 2012.

1. Pawar, S., Dell, A. I. & Savage, V. M. Dimensionality of consumer search space drives trophic interaction strengths. *Nature* **486,** 485–489 (2012).
2. Charnov, E. L., Turner, T. F. & Winemiller, K. O. Reproductive constraints and the evolution of life histories with indeterminate growth. Proc. Natl Acad. Sci. USA, **98,** 9460–9464 (2001).
3. West, G. B., Brown, J. H. & Enquist, B. J. A general model for ontogenetic growth. *Nature* **413,** 628–631 (2001).
4. Kozlowski, J., Czarnoleski, M. & Danko, M. Can optimal resource allocation models explain why ectotherms grow larger in cold? *Integr. Comp. Biol.* **44,** 480–493 (2004).
5. Quince, C., Abrams, P. A., Shuter, B. J. & Lester, N. P. Biphasic growth in fish I: theoretical foundations. *J. Theor. Biol.* **254,** 197–206 (2008).
6. Von Bertalanffy, L. Quantitative laws in metabolism and growth. *Q. Rev. Biol.* **32,** 217–231 (1957).
7. Andersen, K. H., Farnsworth, K. D., Pedersen, M., Gislason, H. & Beyer, J. E. How community ecology links natural mortality, growth, and production of fish populations. *ICES J. Mar. Sci.* **66,** 1978–1984 (2009).
8. Glazier, D. S. Beyond the '3/4-power law': variation in the intra- and interspecific scaling of metabolic rate in animals. *Biol. Rev. Camb. Philos. Soc.* **80,** 611–662 (2005).
9. Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, 1983).
10. Nagy, K. A. Field metabolic-rate and food requirement scaling in mammals and birds. *Ecol. Monogr.* **57,** 111–128 (1987).
11. Woodward, G. *et al.* Body size in ecological networks. *Trends Ecol. Evol.* **20,** 402–409 (2005).
12. Fuiman, L. A. The interplay of ontogeny and scaling in the interactions of fish larvae and their predators. *J. Fish Biol.* **45** (Suppl. A), 55–79 (1994).
13. Abrams, P. A. Functional responses of optimal foragers. *Am. Nat.* **120,** 382–390 (1982).
14. Glazier, D. S. The 3/4-power law is not universal: Evolution of isometric, ontogenetic metabolic scaling in pelagic animals. *Bioscience* **56,** 325–332 (2006).
15. Yodzis, P. & Innes, S. Body size and consumer-resource dynamics. *Am. Nat.* **139,** 1151–1175 (1992).

Author Contributions H.C.G. originated the idea for the paper, carried out the statistical analyses and coordinated the writing of the manuscript. B.J.S., D.T.de K. and P.A.A. helped to write the manuscript and contributed with all discussions that defined the content of this paper. B.J.S. also helped to design Figs 1 and 2.

Competing Financial Interests Declared none.

doi:10.1038/nature11829

Pawar *et al.* reply

REPLYING TO H. C. Giacomini, B. Shuter, D. T. de Kerckhove & P. A. Abrams *Nature* **493,** http://dx.doi.org/10.1038/nature10829 (2012)

Current studies assume that per-capita consumption rates always scale with body mass to an exponent of 0.75. We showed that, contrary to this assumption, consumption rates scale sublinearly (exponent of approximately 0.85) when organisms forage in two dimensions (2D), and superlinearly (exponent of approximately 1.06) when they forage in 3D[1]. Giacomini *et al.* argue that the superlinear scaling in 3D interactions we observed cannot be reconciled with life-history theory for maximal body size[2]. Consequently, they search for biases in our study that might cause this superlinear scaling. However, their comments do not challenge our central result that consumption rates scale superlinearly in 3D, and significantly more steeply than in 2D. We propose instead that life-history theory may need revision to include interaction dimensionality.

The first empirical concern of Giacomini *et al.* is that laboratory studies overestimate consumption rates of larger consumers because container sizes scale sublinearly with consumer size, thus disproportionally reducing predator fear and prey evasion. However, we have already shown that scaling of resource density (abundance per unit-container area or volume) is statistically indistinguishable between 2D (exponent of 0.79 ± 0.09) and 3D (exponent of 0.86 ± 0.06)[1]. As we do not observe a disproportionate increase in resource density in 3D, this argument cannot explain why 3D consumption rates scale more steeply than 2D. Second, Giacomini *et al.* state that our 3D data are biased towards juvenile stages for vertebrates and adult stages for invertebrates. They provide an indirect test (chi-squared test of association) to support this claim and suggest how it might affect our

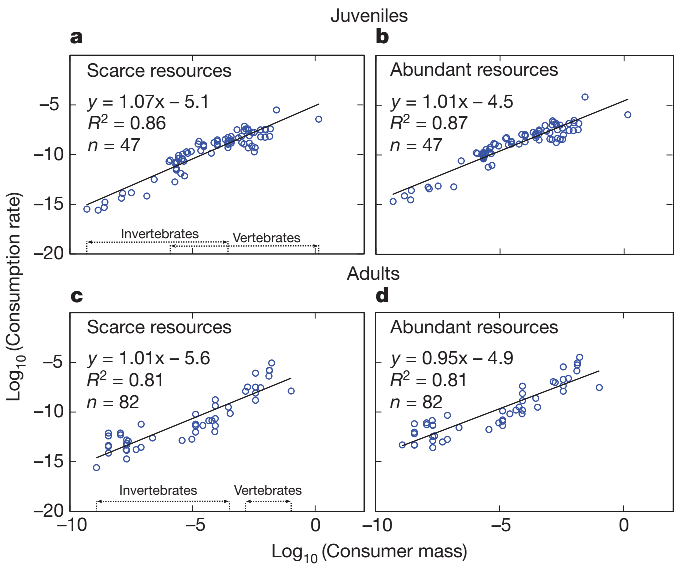


Figure 1 | Effect of ontogeny on scaling of 3D per-capita consumption rates. **a**, **b**, Scaling of juvenile consumption rate (kg s^{-1}) with consumer body mass (kg) at scarce and abundant resource densities. **c**, **d**, Scaling of adult consumption rate. These data are the same as in Fig. 3 of our paper[1], but separated by life stage. Solid black lines were fitted using ordinary least-squares regression. Both scaling exponents (slopes) with scarce resources are superlinear (exponent ± 95% confidence intervals; 1.07 ± 0.10 for juveniles and 1.01 ± 0.15 for adults) and significantly steeper than the 2D exponent of 0.85 (excluded from both confidence intervals), showing that our original results[1] remain unaltered.

results using a schematic (their Fig. 2) that has no relation to values in our data. To test directly for this bias, we analysed our data for juvenile and adult stages separately and found no significant difference in 3D scaling (1.07 ± 0.10 for juveniles and 1.01 ± 0.14 for adults) between juveniles and adults (Fig. 1).

The first theoretical concern of Giacomini *et al.* is that we assume capture success per predator–prey encounter is constant. This claim is incorrect. We assume only that capture success does not vary systematically with body size[3], so our model is consistent with variation in capture success that does not correlate with size. Second, they state that biomechanical constraints can explain why terrestrial organisms foraging in 3D (for example, flying species) are more limited in size than ground dwelling (2D) species, and that our predictions contradict this pattern. This is also incorrect because we make no predictions about maximum body size, and we certainly do not contrast sizes of flying and ground-dwelling species. Maximum body size is a prediction best made by biomechanical theories, whereas our theory makes predictions about the feasible size ratios of consumer–resource pairs. Third, they state that differential scaling for realized and maximal consumption rate implies that organisms above and below certain sizes are not energetically viable. However, their argument is based on the assumption that maximal consumption rate scales as $m^{0.75}$. Our database and analysis are more extensive than those in studies cited[4,5] by Giacomini *et al.*, and indeed a key conclusion of our work is that consumption rates do not scale as $m^{0.75}$, even when there are abundant resources and a maximal consumption rate is expected. To assume that maximal consumption rate scales as $m^{0.75}$ is inconsistent with available data and is counter to our main findings. Attempting to reconcile this assumption with our study will inevitably lead to inconsistent predictions. Fourth, Giacomini *et al.* suggest that superlinear scaling of consumption rate would predict infinitely large body sizes when integrated into life-history models. Apart from the fact that biomechanical and physiological constraints do indeed set strict upper bounds on organismal sizes[6], current life-history theory[7–9] does not incorporate mechanistic models of consumption rate. By revealing a surprising dependence of consumption rate on interaction dimensionality, our study shows why integration of such models is necessary.

Life-history models also cannot account for organisms that shift between 2D and 3D foraging during ontogeny. Such shifts are common[10] and should be considered before combining life-history theory with the superlinear scaling of consumption rate. Moreover, as we emphasized[1], and as Giacomini *et al.* acknowledge[2], our theory is for consumption rate per trophic link, whereas consumers rarely feed exclusively on a single resource. For example, 3D consumers may compensate for the disadvantage of being small through ontogenetic shifts in foraging behaviour or by feeding on multiple resource types. These are interesting areas for future study but do not call into question our original findings[1].

Samraat Pawar[1]†, Anthony I. Dell[1]† & Van M. Savage[1,2,3]
[1]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1766, USA.
e-mail: samraat@uchicago.edu
[2]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA.
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.
†Present addresses: Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA (S.P.); Systemic Conservation Biology, Department of Biology, University of Göttingen, Göttingen, Germany (A.I.D.).

1. Pawar, S., Dell, A. I. & Savage, V. M. Dimensionality of consumer search space drives trophic interaction strengths. *Nature* **486,** 485–489 (2012).
2. Giacomini, H. C., Shuter, B., de Kerckhove D. T. & Abrams, P. A. Does consumption rate scale superlinearly? *Nature* **493,** http://dx.doi.org/10.1038/nature10829 (2012).
3. Savage, V. M. *et al.* Comment on 'The illusion of invariant quantities in life histories'. *Science* **312,** 198 (2006).
4. Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, 1983).
5. Nagy, K. A. Field metabolic rate and food requirement scaling in mammals and birds. *Ecol. Monogr.* **57,** 111–128 (1987).
6. Harrison, J. F., Kaiser, A. & VandenBrooks, J. M. Atmospheric oxygen level and the evolution of insect body size. *Proc. R. Soc. Lond. B* **277,** 1937–1946 (2010).
7. Charnov, E. L. Evolution of mammal life histories. *Evol. Ecol. Res.* **3,** 521–535 (2001).
8. Charnov, E. L. & Gillooly, J. F. Size and temperature in the evolution of fish life histories. *Integr. Comp. Biol.* **44,** 494–497 (2004).
9. Quince, C., Abrams, P., a., Shuter, B. J. & Lester, N. P. Biphasic growth in fish I: theoretical foundations. *J. Theor. Biol.* **254,** 197–206 (2008).
10. Hartvig, M., Andersen, K. H. & Beyer, J. E. Food web framework for size-structured populations. *J. Theor. Biol.* **272,** 113–122 (2011).

doi:10.1038/nature11830

nature OUTLOOK

HEART HEALTH

31 January 2013 / Vol 493 / Issue No 7434

Cover art: Andrew Baker

Despite medical advances, with drugs like statins and devices such as stents, heart disease remains the leading cause of death worldwide. As this Outlook makes clear, only a comprehensive battle plan will defend the heart from a range of modern-life assaults.

Our very surroundings form the battlefield. From poor town planning to energy-saving ergonomics, people face incentives to engage in unhealthy practices (page S4). But as with most medical conditions, the best tactic for preventing cardiovascular disease is to stop it before it starts, and health scientist Joep Perk calls on the heathcare community to pull together (page S6).

Clinicians are trying to predict a person's risk of cardiovascular disease at the earliest possible stage (page S7). Once heart disease is triggered, the best defence requires an understanding of the underlying biochemical processes (page S10). Researchers are revealing ever finer details of cardiovascular disease, down to individual receptors on the surface of cells, as cardiovascular researcher Sébastien Foulquier explains (page S9).

This expanding knowledge of the mechanisms offers hope of many new treatments. For example, atrial fibrillation afflicts tens of millions of people around the world, but computational and chemical tactics can now solve the problem, sometimes within minutes (page S12). Likewise, new drugs promise better treatments for some of the oldest ailments, including high blood pressure and cholesterol (page S14). Researchers are also demonstrating how the mind can trigger cardiovascular disease and perhaps orchestrate effective repairs (page S16). Ultimately, we need a thorough understanding of health and of the heart itself — from single cells to societies — and a range of diverse strategies if we are to defeat heart disease.

We acknowledge the financial support of Bayer in producing this Outlook. As always, *Nature* has full responsibility for all editorial content.

Mike May
Guest Editor

CONTENTS

COLLECTION

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at http://www.nature.com/advertising/resources/pdf/outlook_guidelines.pdf

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, Nature **Vol XXX**, No. XXXX Suppl, Sxx–Sxx (2013). To cite previously published articles from the collection, please use the original citation, which can be found at the start of each article.

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Heart Health* supplement can be found at http://www.nature.com/nature/outlook/heart_health_2013/

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe (excluding Japan): Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea:Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES
Feedback@nature.com

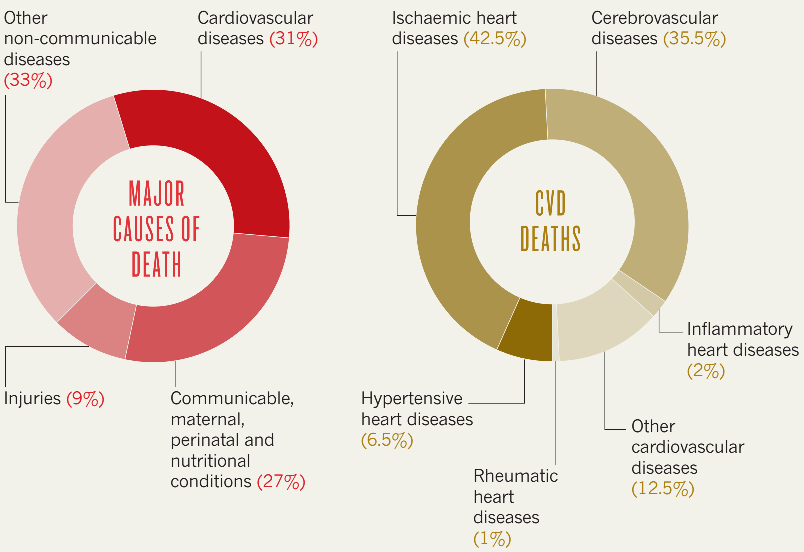
BIOCHEMISTRY TO BEHAVIOUR

Cardiovascular disease (CVD) remains the grim reaper's primary calling card, but people can take steps to keep the world's number one killer at bay.

By **Bill Cannon**.

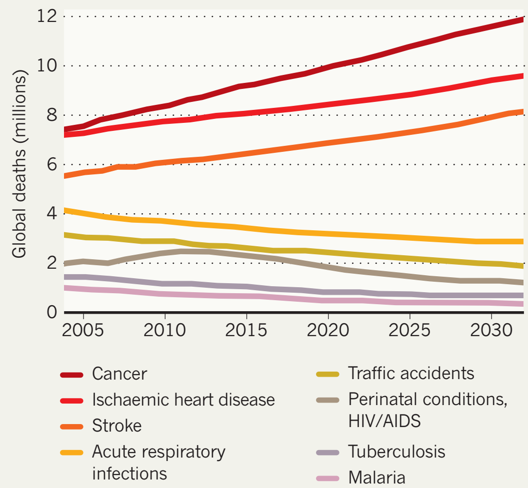
HEARTBREAKING STATISTICS

Cardiovascular disease — pathologies of the heart, blood vessels and the vascular system of the brain — claims more lives than anything else, accounting for nearly one-third of deaths worldwide. Among CVDs, ischaemic heart disease (failure to supply blood to the heart) is caused primarily by clogged arteries (atherosclerosis), it results in the most fatalities among men and women but also represents the largest disparity in heart-disease type between the sexes. Those deaths are projected to increase until at least 2030.



MORE HEART DISEASE AHEAD

Over the next twenty years, significant increases in deaths from coronary diseases — ischaemic heart disease and stroke — will only be outpaced by cancer, while deaths from many other diseases will decrease.

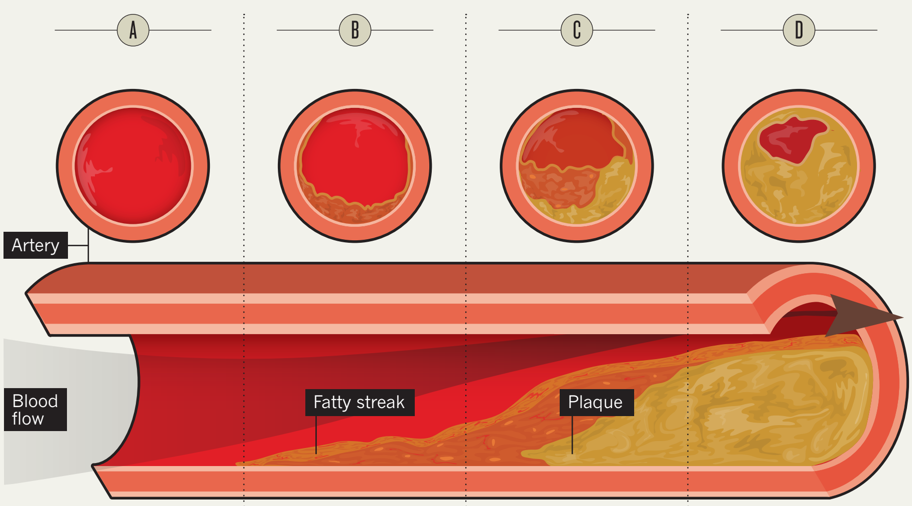


33

global deaths per minute from cardiovascular disease in 2008

WHEN ARTERIES CLOG

Fatty material, cholesterol, cellular waste and other substances can build up in the endothelial cell lining of an artery, leading to plaque formation and blockages that starve the heart of oxygen and cause a heart attack. Cholesterol and triglycerides, high blood pressure and cigarette smoke, which accelerates atherosclerosis, seem to be the leading plaque promoters.



A Healthy arteries allow blood to flow freely, with nothing blocking the channel.

B Atherosclerosis starts with what is known as a 'fatty streak' of material building up in the artery wall.

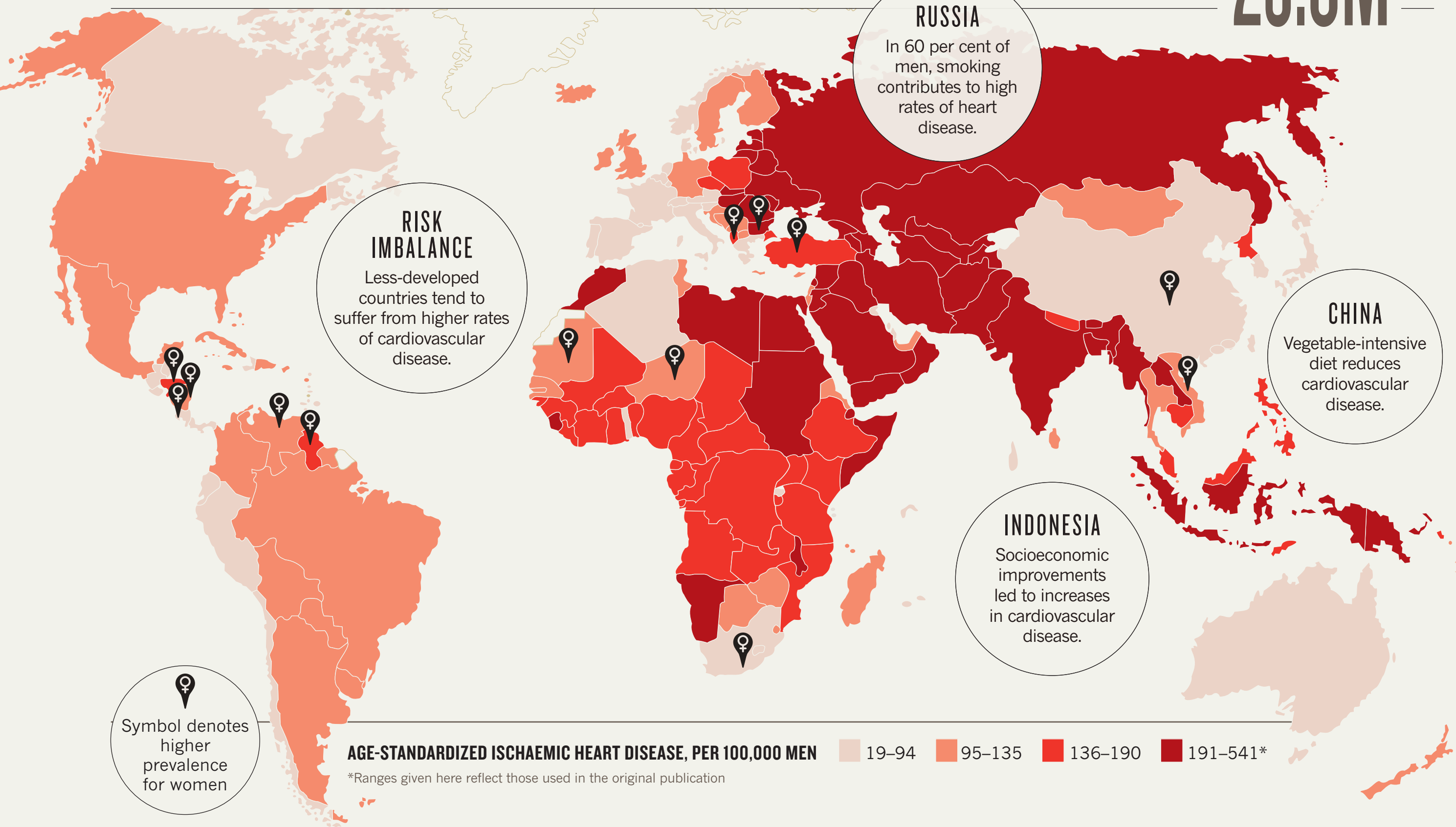
C Over time, the material can form a plaque that might become a thrombosis, or clot.

D A clot might eventually form a potentially fatal blockage in the artery.

GLOBAL HEART-DISEASE BURDEN

The rate of ischaemic heart disease among men varies around the world, and similar variation exists for women.

23.6M the projected annual number of worldwide deaths from cardiovascular disease in 2030



PRINCIPLES OF PREVENTION

The World Heart Federation splits CVD risk into two categories. Members of certain groups — men, the elderly and those with a family history of heart disease — are simply at elevated and 'non-modifiable' risk. Other risk factors shown below can be modified by lifestyle changes.

High blood pressure
Cut blood pressure through weight control, exercise, healthy diet, avoiding tobacco products and reducing alcohol and sodium intake.

Bad body biochemistry
Poor diet, high cholesterol and obesity are closely associated, and so is the remedy: a diet low in fat and high in fruits and vegetables.

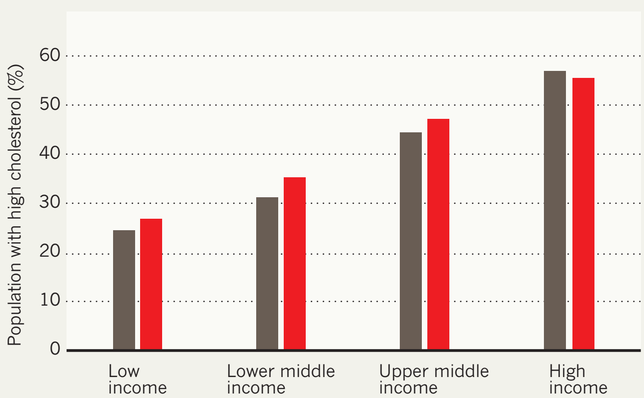
Inactivity
Just 30 minutes of exercise five times a week, or 20 minutes of vigorous excercise three times a week, reduces your risk of cardiovascular disease.

HIGH CHOLESTEROL FOLLOWS HIGH INCOME

People living in high-income countries suffer from high total cholesterol twice as much as people in low-income countries, according to data from the World Health Organization in 2008. So the richest countries might benefit the most from preventive measures, such as improved diet.

Raised cholesterol (≥190 mg dl^{-1}), ages 25+, age adjusted

Male Female



80%

the proportion of all heart disease that might be prevented if people would exercise, eat a healthy diet and not use any tobacco products.

Bill Cannon is a freelance science writer based in Chapel Hill, North Carolina.

COMSTOCK/THINKSTOCK

Freeways in Los Angeles, California, mean that walking across town is not an option, preventing one of the best ways to maintain cardiovascular health.

PUBLIC PLANNING

Designs fit for purpose

Better thought-out town planning and interior design can create healthier environments, but how to effectively implement the best designs remains uncertain.

BY DUNCAN GRAHAM-ROWE

The migration of people to towns and cities in the United Kingdom during the nineteenth century led to squalid conditions and rampant outbreaks of cholera and typhus. Sanitary reformers successfully campaigned for legislation requiring new homes to have running water and adequate drainage. Nearly a century-and-a-half later, we are at a similar juncture: cardiovascular disease is rising at a rate that threatens to bring healthcare services to its knees[1], and urban environments are once again threatening the wellbeing of the people.

Evidence is mounting that the modern lifestyle of working and playing in front of screens — combined with readily available energy-packed food — has people exercising less and eating more[2]. In fact, a recent study concluded that inactivity plays a part in nearly a third of the disability years lived by people with ischaemic heart disease[3]. There are also more subtle forces at work in the way towns and buildings are arranged. Living in suburbs has practically forced people to drive a car to go anywhere, says Gregory Heath, an epidemiologist and public-health scientist at the University of Tennessee at Chattanooga. The positioning of stairs and elevators, the distribution of supermarkets, and the way that suburbs form through unchecked expansion rather than planning can each have a major impact on cardiovascular health.

The long-term cost of sedentary lifestyles will be measured in billions of dollars, says Bengt Kayser, director of the Institute of the Science of Movement and Sport Medicine at the University of Geneva in Switzerland. And yet, says Kayser, if people were to walk for just half an hour each day, the benefits to their health and to national economies would be dramatic. In fact, for those most at risk, walking even 10 minutes a day can improve health[4]. "Physical exercise is like a magic pill," says Kayser.

The problem is that poorly planned urban landscapes discourage exercise and healthy eating. Can a twenty-first-century revamp of the urban environment encourage people to live more active, healthy lives?

PUSHING PEOPLE

One of the biggest contributors to this problem is one of ergonomics. The office desk may seem less sinister than the noxious substances and mechanical looms of nineteenth-century factories. Nevertheless, sedentary lifestyles are taking a toll on human health. Getting people to take the less easy option — such as taking the stairs instead of the lift — can make a huge difference. Although providing alternatives to stairs is necessary for disabled access, buildings should be designed so that stairs are a conspicuous and attractive option for those able to use them. The bouts of exercise involved in climbing a set of stairs are a boon to cardiovascular health.

Unfortunately, Kayser says, offices and public spaces often encourage people to take the easier option. The lifts are usually in a prominent position, whereas access to the stairwell is tucked away. And lifts tend to be placed right next to the stairs, offering an easy ride. Still, Kayser says, many people will take the stairs if that is the shortest route. One approach being tried is to make stairs more fun with posts of humorous or encouraging messages, or by designing the steps to be interactive, such as the piano-like steps leading out of the Odenplan metro station in Stockholm: climbing up and down the steps makes music.

Another tactic is to implement policies that encourage commuters to use public transport,

a mode of travel that tends to involve more walking and cycling than does the car, says David Ogilvie, co-investigator in the Centre for Diet and Activity Research at the University of Cambridge, UK. Possible ways to achieve this are by making public transport more attractive through the combined use of congestion charges on cars — an automatic charge when a vehicle enters the city centre, as happens in London — and the introduction of exclusive lanes, paths and trails for cyclists.

Indeed, a comparison by Heath's team of two poor neighbourhoods in Tennessee showed that young people with access to a two-mile length of extra-wide urban trail that could accommodate both cyclists and pedestrians were nearly twice as likely to be physically active than those living in the neighbourhood lacking such an amenity[5]. But Heath stresses that town planners shouldn't expect such results just by slapping down a low-quality bike path. The Tennessee trail, he says, was designed to be safe, aesthetically pleasing and well lit to discourage crime, while connecting local communities with nearby schools, a recreational centre, library and shops. It's also well maintained to stop foliage limiting visibility. Rest areas and open space along the route "translates into more eyes on the 'street' path, which deters crime and threats to safety", says Heath.

Ogilvie agrees, pointing out that there are countless examples of poorly thought-out cycle lanes that do little to promote cycling —including some that have poor visibility, an ambiguous right-of-way or are located too close to parking spaces, with the risk of car doors taking out a passing cyclist. Rather than promoting cycling, people are being put off. "We already know a lot about the benefits of physical activity to health," Ogilvie says, "but we know less about the effects of the environment on physical activity." What's more, he and colleagues have reported evidence of such measures leading to more active lifestyles, yet the improvements were modest, with people making on average only eight additional cycle trips per year[6]. That's to be expected, says Ogilvie, in part because people tend to stick to travelling patterns.

IS BIGGER BETTER?

In 2007, the Australian government issued guidelines for 'liveable neighbourhoods' to encourage physical activity as a part of daily routines. "What's happening in Australia, and to some extent in the US, is people are looking for affordable housing so they look on the fringes of cities," says Billie Giles-Corti, a social epidemiologist at the Melbourne University School of Population Health. The problem is that people are then forced to drive everywhere because there is not yet any infrastructure such as public transport and local amenities.

To address this concern, the state government of Western Australia issued planning guidelines aimed at creating urban environments that encouraged walking, cycling and public transport. "What you really want is to increase [population] density," says Giles-Corti. As long as density is low, she says, public transport and other local services will be poor and residents will have to drive more. Only 17% of Australians walk enough for it to benefit their health — a statistic that hints at opportunities for public-health improvement, says Giles-Corti. "Active transport is generally habitual while recreational walking is volitional." It's better, she says, to "get people active as part of their day rather than simply a recreational activity".

One tactic might be for policymakers to incentivize businesses and transport companies to extend networks to low-density areas on the outskirts. Based on current trends they would be sound investments because population densities are only going to increase, says Giles-Corti. "By 2050 as much as 70% of the world's population will be living in cities," says Giles-Corti.

K.J.VOGELIUS/GETTY IMAGES

Musical stairs: a chance to tiptoe a tune attracts walkers at Stockholm's Odenplan subway station.

THE FOOD FACTOR

Physical activity is only part of the story of how buildings and infrastructure affect health. Charles Abraham, professor of behavioural change at the University of Exeter Medical School, UK, argues that we should be putting more emphasis on understanding the energy content of the food our environment makes available. Of particular concern, says Heath, is the emergence of vast urban areas, with low-income residents, that lack traditional shops or supermarkets. According to a 2009 report by the US Department of Agriculture's Economic Research Service, as many as 11.5 million low-income people in the United States — about 4% of the country's population — live more than 1.6 kilometres from a fully stocked supermarket as opposed to a convenience store[7]. Such distances severely limit access to healthy food, leaving residents on a diet of junk food found in local shops. These risk factors for cardiovascular disease could be avoided through planning policies that encourage urban locations for farmers' markets, community gardens and even mobile markets, Heath says.

Furthermore, billions of dollars are spent each year trying to persuade people to buy unhealthy foods — and this barrage of advertising often works. Brian Wansink at Cornell University in New York has made a career out of highlighting what he calls "mindless eating". Wansink studies how supermarket layout, food packaging and even the way food is presented on a plate can influence diet. His research group recently demonstrated that sales of healthier food could be boosted in a school canteen by making healthier options such as salads or fruit more convenient to buy[8].

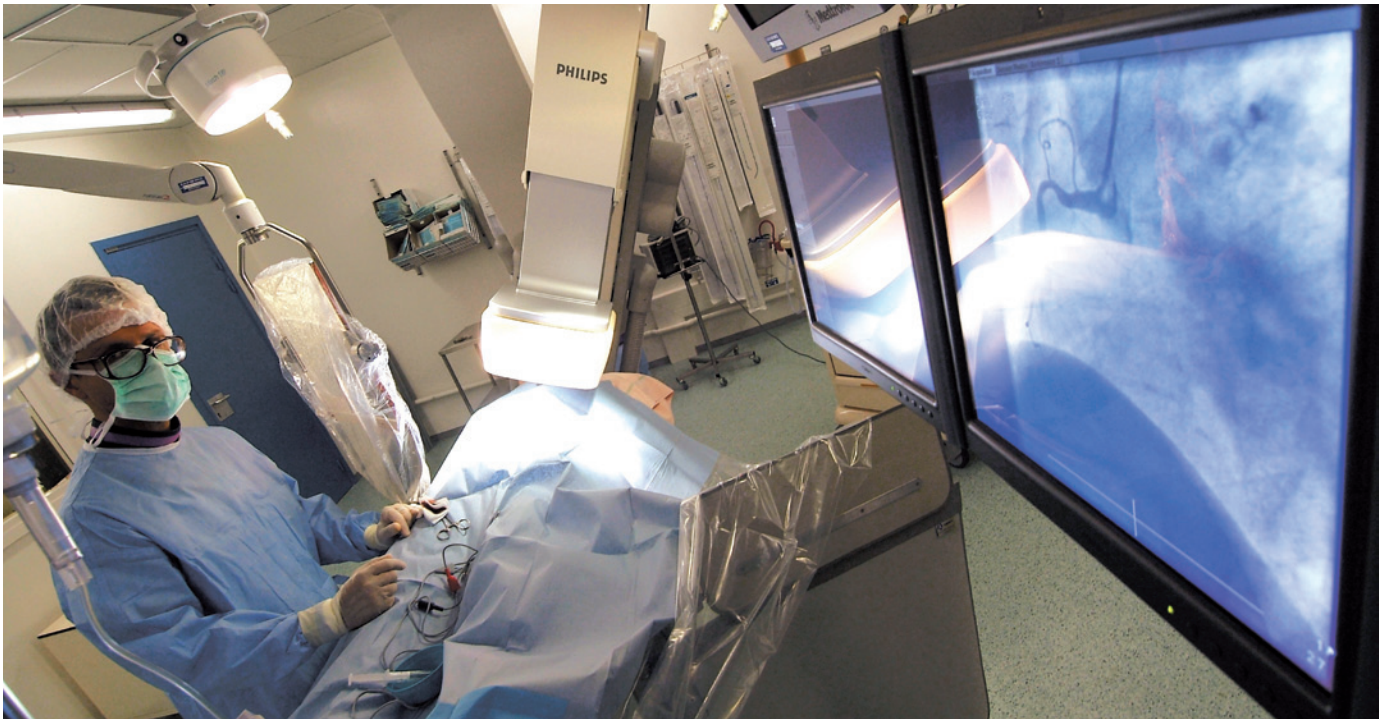
Even for those seeking healthier food it can be a minefield out there. Better labelling can make it obvious how healthy a product is based on its fat, salt, saturated fats and sugar content. Foods high in sugar and saturated fats could also be taxed higher to help fund public-health services. "These measures would greatly reduce consumption of unhealthy food without preventing food and drink manufacturers from making a profit," says Abraham.

What worries health policymakers most are the hard-to-reach populations. Even in cases where policy leads to some improvements, to what extent will these reduce cardiovascular disease in the people who are least active? Most studies to date have not identified who benefits most from intervention, says Ogilvie. Is it those most at risk, or people who were already fairly active in the first place and needed little encouragement to increase their activity levels? That, says Ogilvie, is the important question. ■

Duncan Graham-Rowe *is a freelance science writer based in Piltdown, UK.*

1. *Political declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-communicable Diseases* (UN draft resolution, 2011).
2. Myers, J. *et al. N. Engl. J. Med.* **346,** 793–801 (2002).
3. Lim, S. S. *et al. Lancet* **380,** 2224–2260 (2012).
4. Kayser, B. *et al. Eur. J. Cardiovasc. Prev. Rehabil.* **17,** 569–575 (2010).
5. Heath, G. W. *et al. J. Physical Activity and Health* **3** (Suppl. 1), S55–S71 (2006).
6. Yang, L. *et al. Br. Med. J.* **341,** c5293 (2010).
7. *Access to Affordable and Nutritious Food: Measuring and Understanding Food Deserts and Their Consequences* (US Department of Agriculture, 2009).
8. Hanks, A. S. *et al. J. Public Health* **34,** 370–376 (2012).

GABRIELLE VOINOT/LOOK AT SCIENCES/SCIENCE PHOTO LIBRARY



Coronary angiography uses a contrast agent and X-ray imaging to search for blockages in the heart's blood vessels.

DIAGNOSTICS

The new risk predictors

New imaging methods and biomarkers may help identify people who are at risk for heart disease but are overlooked by standard risk assessments.

BY PETER GWYNNE

Doctors have long recognized the basic risk factors for cardiovascular disease: age, sex, high blood pressure, high cholesterol, smoking, diabetes and a family history. But these criteria have two significant limitations: many people with all these risk factors do not suffer heart problems, and dying of a heart attack is hardly unknown in people with none of the risk factors. According to cardiologist Jonathan Murrow at Georgia Health Sciences University in Augusta, one-third of the sudden deaths arising from coronary artery disease come without warning[1].

Clinicians have developed new ways of calculating the risk of cardiovascular disease (CVD) for those without any of the traditional risk factors. One new approach involves screening patients for certain biomarkers — through blood tests or imaging technology — that correlate with a higher than usual risk of a major cardiovascular event. So far, the cardiology community remains divided on the value of such non-traditional methods.

Supporters of the traditional approach argue that the new risk predictors aren't much better. Beyond that, new methods can potentially cause harm. In the view of Mark Ebell, an epidemiologist at the University of Georgia in Athens, even relatively innocuous blood tests can lead to over-diagnosis and over-treatment and imaging studies have the added danger of radiation exposure[2].

Advocates of using biomarkers and imaging technology don't expect the new methods to replace the old. Rather, they see them improving the outcomes in some individuals and groups of people. For example, improved risk prediction could personalize disease prevention depending on lipid levels or blood pressure. "There are niche populations for whom biomarkers can be helpful," says John Wilkins, who specializes in cardiology and preventive medicine at Northwestern University Feinberg School of Medicine in Chicago, Illinois.

One of the populations that could benefit most is women. "About a third of women who have had a heart attack don't have significant narrowing of the coronary arteries," explains Martha Gulati, who heads the Preventive Cardiology and Women's Cardiovascular Health section at Ohio State University in Columbus. "So now we are looking for the small blood vessels in women." That requires new tests, most notably a cardiac form of magnetic resonance imaging (MRI).

➲ NATURE.COM
Discussing long-term risks of heart disease with patients:
go.nature.com/xlnbkh

RANKING RISKS

Cardiovascular disease comes in several forms, including problems with blood vessels and heart valves, arrhythmias (irregular heart beat), heart attacks and strokes. Predicting the risk varies with the form of CVD. Coronary artery disease, for example, is more predictable. "The most widely used coronary heart disease risk prediction that we use in our current cholesterol-lowering guidelines is the Framingham risk score," Wilkins says. "It is the most strongly validated and robust for 10-year estimates." The score is based on data from the Framingham heart study, a 64-year-old project that aims to identify the common factors or characteristics that contribute to CVD by monitoring the health of various cohorts of people.

The Framingham model involves knowing an individual's age, sex, systolic blood

pressure, total cholesterol level, high-density lipoprotein (HDL) cholesterol level, smoking status and the presence or absence of medication to treat high blood pressure, explains Gregg Fonarow, director of the Ahmanson-UCLA Cardiomyopathy Center in Los Angeles, California. Given those details, the model computes the probability that the individual will suffer a CVD event in the next 10 years, generally segregated into either high risk, intermediate risk or low risk. The predictions provide more than peace of mind for some patients, says physiologist Kerry McDonald at the University of Missouri School of Medicine in Columbia. They also generate the data needed to design a treatment regime appropriate to the level of risk.

The realization that a low risk of CVD in the short term can conceal a much higher risk throughout a lifetime has added to the argument for more sensitive ways to access risk. A study headed by Donald Lloyd-Jones, a cardiologist at Northwestern University Feinberg School of Medicine in Chicago, Illinois, measured the CVD risk factors of more than 250,000 individuals at the ages of 45, 55, 65 and 75 (ref. 3).

"The 10-year risk equations do a reasonable job," Lloyd-Jones explains. "However, they will give low-risk estimates even to young people with elevated risk factors, because they are young." But any elevated risk factor can have long-term consequences. For example, a 45-year-old man whose risk factors are all optimal has only a 1.4% chance of a heart attack, stroke or other form of fatal heart condition in his lifetime; but two or more risk factors at that age increase the risk dramatically, to 49.5%. Thus Lloyd-Jones's team and others have developed algorithms for 30-year and lifetime risks. "These are at the same stage of development as 10-year risk assessments were 10 years ago," Lloyd-Jones says. "People are starting to use them to see if they motivate patients with elevated risk factors to change their lifestyles or adhere to their therapies."

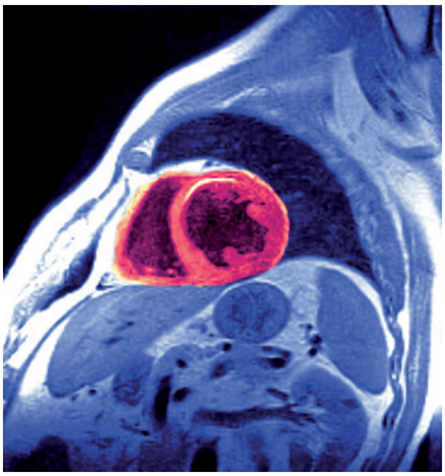
Almost all doctors now realize the importance of emphasizing to young adults the importance of a healthy lifestyle — including diet, regular exercise and not smoking. And cardiologists have realized that some members of the Framingham intermediate group could benefit from drug treatments rather than just lifestyle changes[4]. Those factors suggest that certain patients can benefit from predictive methods more detailed than the Framingham risk score.

PROSPECTS FOR PREDICTION

In mid-2012, one new insight seemed so obvious as to be trivial. Two twenty-year studies headed by nutritionist Lu Qi of the Harvard School of Public Health in Boston, Massachusetts, revealed that individuals with blood types A, B and AB face a higher risk of coronary heart disease than the 44% of Americans with blood type O. People with the rare AB blood type have a heart disease risk 23% higher than O types, and those with types B and A have 11% and 5% increased risk, respectively[5]. Qi believes that blood typing provides another tool for predicting who is at risk of CVD.

Other predictive techniques involve medical imaging. Among these, scanning for coronary-artery calcium holds the most promise. "In the right patients, it seems to be the best risk predictor to use in conjunction with the Framingham risk score," Wilkins says. "However, several other serum and imaging-based biomarkers may assist in risk prediction in specific patient populations as well."

The calcium scan uses computed tomography (CT) to detect a build-up of calcium in the coronary arteries, a possible indicator of impending atherosclerosis and susceptibility to heart attack. One study looked at the possible value of coronary artery calcium scores (CACS)[6]. About a quarter of the people who would have been designated as intermediate risk according to Framingham risk score methods were determined to be high risk when their CACS were taken into account.



This MRI scan through the chest (blue) reveals dangerous fat (yellow) in the walls of the heart (red).

SOVEREIGN, ISM/SCIENCE PHOTO LIBRARY

Fonarow points to one disadvantage of the procedure: CT scans expose patients to potentially harmful ionizing radiation. "Some studies suggest screening for coronary calcium in intermediate-risk patients might be useful," he says. "But is the radiation risk worth it?"

That question remains to be answered. However, new techniques such as ultrafast CT minimize the radiation dose, says Gulati. And several scanning techniques applicable to CVD do not involve ionizing radiation. "A lot of echocardiography is pretty non-invasive," says McDonald. "There's a movement toward non-invasive tests that don't exacerbate the problem."

NEW AND NON-INVASIVE

Two non-invasive imaging methods have shown promise in women. The methods have particular value as the differences between women and men's medical history become increasingly evident. "Women aren't small men," Gulati says. "We're physiologically different." Thus women who smoke have a greater risk of CVD than men who smoke, and diabetic women face three times greater chance of getting heart disease than men with diabetes.

To study small blood vessels in women, Ohio State University's cardiology department is developing a cardiac MRI. "Rather than just running chemicals through a patient's coronary arteries, we can physically stress them on a treadmill under magnetic resonance imaging," Gulati explains. "This appears to be far more useful to detect microvascular disease in women."

Another technique with particular value for women is intravascular ultrasound. "This can show diffuse plaque along the entire artery rather than specific blockages," Gulati says. "Some medical centres will do an intravascular ultrasound if they see evidence of heart attacks in normal arteries."

Studying biomarkers to improve CVD risk prediction has produced fewer encouraging results. A report by the Framingham heart study in 2006 compared the effects of several novel biomarkers with the traditional risk factors. The biomarkers included several types of protein: natriuretic peptides (protein hormones secreted by heart cells); C-reactive proteins (blood proteins linked with inflammation); fibrinogen (a protein involved in coagulation); urinary albumin (a protein from the kidneys) and the amino acid homocysteine. The report also studied the 'multimarker' approach that combines the predictions of several biomarkers[7]. The study concluded that using 10 biomarkers adds only moderately to the ability to assess risk. As Fonarow sees it, that comment applies to the entirety of alternative methods. "Thousands of studies touting genetic risk factors, gene polymorphisms, biomarkers and a variety of invasive and non-invasive imaging tests have generally not improved on the standard risk model, or the improvement has been very modest — not enough to be cost effective or helpful for effective practice," he says. "Some may be considered in intermediate risk patients, but do they improve clinical outcome?"

Instead, Fonarow wants clinicians to wring more value from existing tools. "The real challenge in clinical practice is that the tried and true Framingham risk model is often not applied," Fonarow says. "The issue is applying the model into clinical practice and getting patients to adhere to therapies." ■

Peter Gwynne *is a freelance science writer based in Sandwich, Massachusetts.*

1. Murrow, J. R. *Am. Fam. Physician* **86,** 398–401 (2012).
2. Ebell, M. H. *Am. Fam. Physician* **86,** 405–406 (2012).
3. Berry, J. D. *N. Engl. J. Med.* **366,** 321–329 (2012).
4. Yeboah, J. *et al. J. Am. Med. Assoc.* **308,** 788–795 (2012).
5. Qi, L. *Arterioscler. Thromb. Vasc. Biol.* **32,** 2314–2320 (2012).
6. Polonsky, T. S. *et al. J. Am. Med. Assoc.* **303,** 1610–1616 (2010).
7. Wang, T. J *et al. N. Engl. J. Med.* **355,** 2631–2639 (2006).

PERSPECTIVE

A tale of two receptors

A hormone system adapted for self-preservation can break and fix your heart, say **Sébastien Foulquier**, **Ulrike Muscha Steckelings** and **Thomas Unger**.

Imagine someone cut severely in an accident and losing blood fast. Blood pressure drops dangerously low as blood is lost, and the kidneys react by releasing the enzyme renin. The circulating renin leads to the production of angiotensin II, a hormone that constricts blood vessels and prompts the kidneys to retain more salt and water to prop up the blood pressure. This renin–angiotensin system can save the life of our imagined person as they lay bleeding to death. Under other circumstances, it can kill. Our research reveals that the renin–angiotensin system might be tweaked in novel ways to fight heart disease — but we can only do that accurately and precisely by first learning the exact details of how the system's hormone receptors work.

The consequences of activating the renin–angiotensin system depend on the circumstances, and to which receptor the angiotensin II binds. If it binds to the angiotensin AT_1 receptor, the blood pressure-increasing mechanism is engaged. And AT_1 receptor binding under normal circumstances can raise blood pressure and damage the cells of the heart's left ventricle, the chamber that pumps oxygenated blood out of the heart. In addition, clinical studies have shown that drugs that block this receptor can also reduce damage to the heart. Therefore, blocking this receptor serves as a therapeutic strategy for treating cardiovascular diseases, especially hypertension.

THE SAME MECHANISM THAT HELPS RESPOND TO HAEMORRHAGE CAN ALSO, UNDER DIFFERENT CIRCUMSTANCES, DAMAGE THE HEART.

When drugs for high blood pressure block a receptor docking site for angiotensin II, more of the hormone remains free to circulate in the blood. Increased levels of circulating angiotensin II affect a second receptor, AT_2. The AT_2 receptor triggers a molecular cascade that expands the size of blood vessels and helps prevent damage to the heart. In essence, the actions triggered by the second receptor are the opposite of the processes caused by the first. In addition, the body makes more AT_2 receptors in response to several pathological conditions, including heart attack, heart failure and stroke. So when things go wrong with the cardiac system, the AT_2 receptor system appears to serve a protective function.

Other research results support such a protective role for the AT_2 receptor. In one study, for example, researchers engineered animals with varying amounts of AT_2 receptor and confirmed that it can activate mechanisms that protect the heart[1] (some studies produced conflicting results, possibly due to different receptor production levels in the genetically altered animals). As well as modifying the amount of the receptor, researchers can look for molecules other than angiotensin II that can turn on the AT_2 receptor. Just such a molecule exists, and it's called compound 21 (C21).

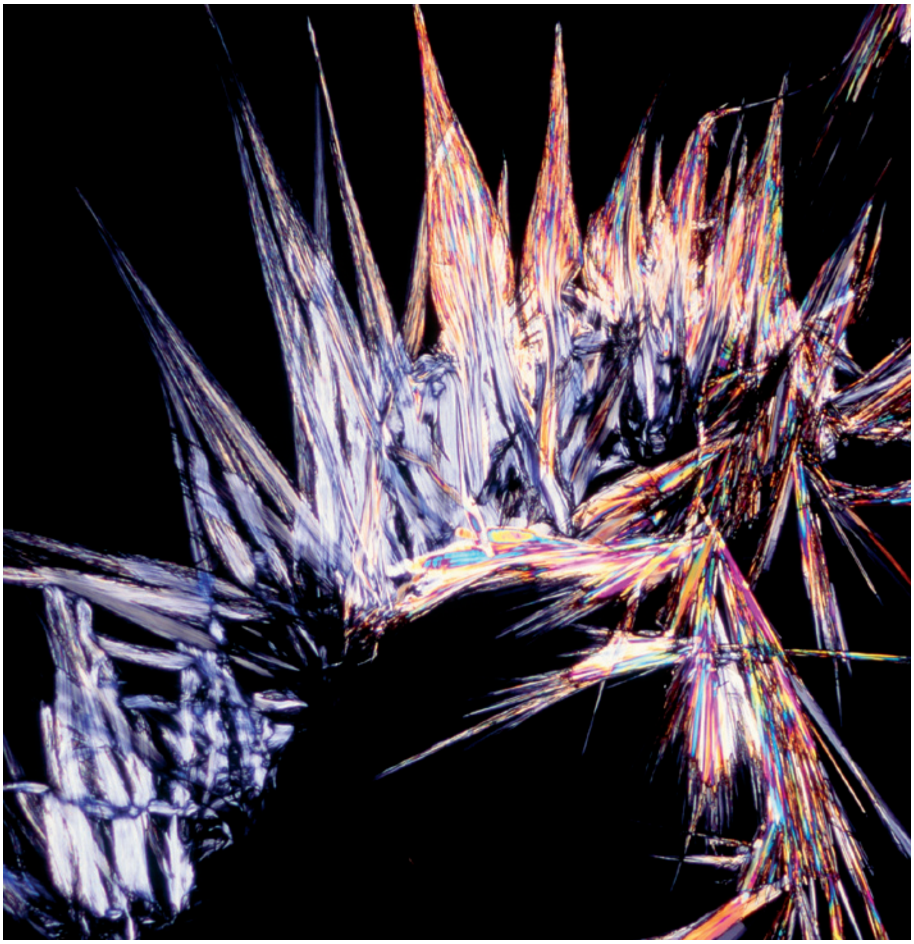
We and other research groups have tested C21's ability to influence high blood pressure, but surprisingly C21 has no effect on the condition[2]. Nonetheless, our results continue to suggest the potential of exploiting this compound's role in repairing an injured heart. For example, C21 reduces stiffness in arteries, an effect that could indirectly affect blood pressure[3,4]. Moreover, in an animal model of a heart attack, in which blood flow to the heart is stopped, causing muscle damage or even death, administered C21 improves cardiac function and reduces scarring to cardiac tissue[5]. These benefits arise from C21's anti-inflammatory properties and its canny ability to keep cardiac cells from dying. And both result from the compound binding to AT_2 receptors. Other findings suggest that C21 can improve heart health in even more ways. In human cardiac stem cells, for example, stimulating the AT_2 receptors slows the death of heart muscle cells[6]. Likewise, activating AT_2 receptors bound to immune cells in specific parts of the heart makes these cells fight inflammation, thereby protecting the heart[7].

The renin–angiotensin system illustrates the complexity of molecular pathways that can help and hinder the heart. The same mechanism that helps respond to haemorrhage can also, under different circumstances, damage the heart. Conversely, using this same system to trigger a different pathway — through a different receptor — repairs heart damage. As scientists explore this pathway through C21, we keep learning more about the molecular mechanisms behind this complex system. Perhaps combining conventional medication for high blood pressure with drugs developed around the AT_2 receptor will provide us with new ways to preserve heart function. As we also now know, even more pathways originate in the renin–angiotensin system, pathways that may lead to future heart medications. But first, we must understand the receptors that mediate it all. ■

Sébastien Foulquier *is postdoctoral researcher at CARIM–School for Cardiovascular Diseases at Maastricht University, The Netherlands. email: s.foulquier@maastrichtuniversity.nl*
Ulrike Muscha Steckelings *is associate professor at the Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark.*
Thomas Unger *is scientific director of CARIM.*

1. Steckelings, U. M. *et al. J. Hypertens.* **28** (Suppl. 1), S50–S55 (2010).
2. Foulquier, S., Steckelings, U. M. & Unger, T. *Curr. Hypertens. Rep.* **14,** 403–409 (2012).
3. Paulis, L. *et al. Hypertension* **59,** 485–492 (2012).
4. Rehman, A. *et al. Hypertension* **59,** 291–299 (2012).
5. Kaschina, E. *et al. Circulation* **118,** 2523–2532 (2008).
6. Altarche-Xifró, W. *et al. Stem Cells* **27,** 2488–2497 (2009).
7. Curato, C. *et al. J. Immunol.* **185,** 6286–6293 (2010).

Competing financial interest
Ulrike Muscha Steckelings declares a conflict of interest. She received research support from Vicore Pharma, which developed compound 21.



SIDNEY MOULDS/SCIENCE PHOTO LIBRARY

Cholesterol crystals accumulate inside plaques — a build-up causes cardiovascular disease.

PATHOLOGY

At the heart of the problem

Research is illuminating the molecular mechanisms that can cascade into debilitating heart disease.

BY CASSANDRA WILLYARD

After the end of the Second World War, the number of deaths related to heart disease in the United States skyrocketed. At that time, the causes of heart disease weren't well understood. So in 1948, the US Public Health Service launched a landmark study in Framingham, Massachusetts, and recruited more than 5,000 of its townspeople. The Framingham Heart Study helped identify many of the major risk factors: smoking, high cholesterol, high blood pressure, diabetes, obesity and lack of exercise. They also identified factors that reduce the risk of heart disease.

In recent decades, researchers have begun untangling the molecular pathways that underlie these risk factors. Even seemingly straightforward relationships, such as the link between high cholesterol and heart disease, are turning out to be more complex than previously believed, which has researchers questioning even the most basic assumptions.

Heart disease is an umbrella term that includes a wide variety of ailments. By far the most common, however, is coronary heart disease, which occurs when fatty plaque builds up in the arteries that feed the heart — a process called atherosclerosis. Plaque can restrict blood flow, but the bigger problem occurs when a plaque ruptures, spilling its contents into the bloodstream and causing a clot to form. A clot that cuts off blood flow to the heart will cause a heart attack. If the clot blocks flow to the brain, the result is a stroke.

Cholesterol is a major component of plaque. "The body itself cannot break down cholesterol. It has to store it or export it, or turn it into something else, like a hormone," says Frank Sacks, a specialist in cardiovascular disease prevention at Harvard School of Public Health in Boston, Massachusetts. "So if cholesterol goes into an artery, it's stuck there." Lipoproteins (conglomerations of lipids and proteins) shuttle cholesterol through the body. A surplus of low-density lipoprotein (LDL), often called 'bad cholesterol', can lodge in artery walls, promoting the formation of plaque and increasing the risk of a heart attack or stroke.

THE COMPLEX ROLE OF HDL

High-density lipoprotein (HDL), a larger molecule, clears away cholesterol-laden plaque from the artery walls and carries it to the liver for disposal. Epidemiological studies consistently show that people who have high levels of HDL cholesterol in their blood have fewer heart attacks. More recent studies, however, suggest that HDL might not be universally protective. "You lower LDL, you reduce heart disease any way you do it," Sacks says. But raising HDL is not always beneficial. "We know there's something protective about HDL," Sacks says. "We cannot really identify what it is."

Part of the problem is that HDL is a mix of molecules that differ in size and composition. Sacks and his colleagues speculated that some types of HDL might be more protective than others. Their previous work pointed them to a small protein found on the surface of some lipoproteins called apolipoprotein (apo) C-III (ref. 1). Studies show that LDL bearing apoC-III promotes plaque build-up in the arteries. Sacks wondered what effect the protein might have on the function of HDL.

So Sacks and his colleagues examined blood samples collected as part of two large epidemiological studies: the Nurses' Health Study, which included about 122,000 female nurses, and the Health Professionals Follow-up Study, which included about 52,000 men. Not surprisingly, Sacks found that HDL was protective; individuals who had higher levels had fewer heart attacks[2]. However, when he focused on HDL with apoC-III, which made up about 10–15% of total HDL, he uncovered an adverse effect in both study groups. When they pooled the two studies, evidence for the adverse effect grew even stronger. HDL with apoC-III "actually predicted a higher, not a lower, rate of heart disease", Sacks says. When the team excluded this harmful HDL, the protective effect of the remaining HDL grew even stronger.

If HDL were directly involved in protecting against heart disease, then individuals who

have genetic variants associated with increased HDL production should have fewer heart attacks. Indeed, variants that lower LDL are consistently associated with a reduced risk of heart attacks. But the reverse doesn't seem to hold true for HDL. A group of researchers examined the effect of a genetic variant that boosts HDL cholesterol levels, but individuals with this variant did not have fewer heart attacks[3].

These and other studies suggests that "we still need to learn more about HDL biology and recognize that it's a complex molecule in order to be sure that we develop the best therapeutic strategy," says Gary Gibbons, director of the US National Heart, Lung, and Blood Institute in Bethesda, Maryland.

AN INFLAMED HEART

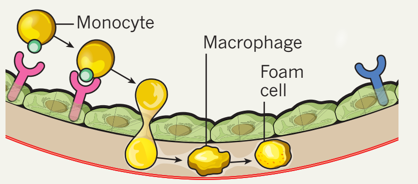
The mechanisms that lead to atherosclerosis are just as complex as the mechanisms that protect against it. For example, research over the past three decades suggests that inflammation plays an intermediary role.

In the 1980s, scientists found that plaque contains macrophages — immune cells that play a role in inflammation. Researchers now suspect that the presence of macrophages results from a cascade of events. The LDL in artery walls prompts endothelial cells, which line the vessels, to produce sticky molecules that snag macrophage precursor cells, monocytes, from the blood. At the same time, endothelial and smooth muscle cells begin to pump out chemicals that attract monocytes. These newly recruited monocytes enter the arterial wall and mature into macrophages, which gorge on LDL cholesterol and balloon in size. "You can't have atherosclerotic plaque without having cholesterol-laden macrophages in your artery wall. That's the major cell type that accumulates," explains Stanley Hazen, an endocrinologist who is head of preventive cardiology and cardiac rehabilitation at the Cleveland Clinic in Ohio. Because their interior is dotted with fatty globs of cholesterol, pathologists call them foam cells. These foam cells are the hallmark of atherosclerotic plaque; they and other immune cells produce chemical signals that prompt smooth muscle cells to migrate to the top of the plaque and form a tough cap. As foam cells accumulate, some die. The cap keeps this pool of living and dead foam cells safely walled off from the blood stream.

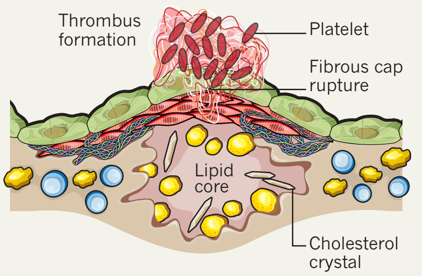
Inflammation seems to increase the likelihood of plaque rupture, which sparks the formation of potentially lethal clots. Collagen and other materials make the cap strong. But inflammatory mediators in the plaque can inhibit collagen synthesis and ramp up the production of enzymes that break down collagen, leading to a thin and weakened fibrous cap, says Peter Libby, chief of cardiovascular medicine at Brigham and Women's Hospital in Boston. Libby's lab and other research groups have also shown that inflammatory mediators can prompt macrophages in the plaque to churn out tissue factor, a protein that promotes clotting. "So not only is inflammation involved in the very first steps of atherosclerosis, but also in the ultimate complications," Libby says.

PROBLEMS FROM PLAQUE

Various cells and cholesterol can interact to block arteries, which is a major cause of heart disease.



1 Blockage begins when the arterial surface captures immune cells (monocytes), which enter the wall and become macrophages that turn into foam cells.



2 The foam cells and cholesterol crystals collect in a lipid core. A rupture in the arterial wall releases the material in the plaque, which reduces blood flow in the artery.

Libby recently added another subplot to this story[4]. Using mice genetically predisposed to develop plaque in their arteries, he and his colleagues showed that a heart attack worsens atherosclerosis by boosting the release of stem cells from the bone marrow. These cells travel to the spleen and become monocytes. The researchers propose that the anxiety and pain associated with a heart attack triggers the sympathetic nervous system to boost stem-cell release and that interrupting this chain might help to prevent future heart attacks. The team also found high numbers of stem cells in the spleens of patients who had died of a heart attack, which suggests the process might work in a similar way in humans[5].

MICROBIAL INFLUENCES

Some of the factors involved in the development of atherosclerosis are entirely new and unexpected. Hazen stumbled across one of these surprising mechanistic pathways a few years ago while looking for metabolites linked to atherosclerosis. Hazen identified a group of compounds "strikingly associated with cardiovascular risk", he says. Some were metabolites that could only be produced by bacteria. "Once we saw that bacteria were likely involved in the pathway, we started looking to the gut," Hazen says. His focus turned to one compound: trimethylamine *N*-oxide (TMAO).

Gut microbes do not produce TMAO directly. They convert phosphatidylcholine (a common component of animal products such as meat and eggs) into a foul-smelling gas called trimethylamine; the liver then converts this gas into TMAO. When Hazen's team gave the mice TMAO, "that alone was sufficient to accelerate atherosclerosis", Hazen says. Even the trimethylamine precursor promoted atherosclerosis when the TMAO-producing microbes were present. But the researchers found they could protect against this by killing the mice's gut bacteria with antibiotics.

There are various reasons why TMAO might promote atherosclerosis. Hazen and colleagues found that the metabolite increases the number of receptors on the surface of macrophages that bind to LDL, which makes the cells more prone to gobble up cholesterol[6]. "This pathway sits right at the junction between cholesterol metabolism and inflammation," Hazen says. "It's influencing both in the artery wall."

If microbes are part of the problem, could antibiotics be part of the solution? Hazen points out that a number of randomized controlled trials have tested whether antibiotics can prevent heart disease, but none proved fruitful. The problem might have been that the drugs tested failed to wipe out the TMAO-producing organisms. Microbes can quickly develop resistance. "We do all our mouse studies now not with a single antibiotic, but with a big gorilla cocktail of five different antibiotics," Hazen says. But he doesn't advocate antibiotic cocktails for the prevention of heart disease. Instead he envisages using probiotics — beneficial microorganisms — to promote healthy microflora, or drugs to interrupt the pathway without killing the bacteria.

Much has changed since the launch of the Framingham Heart Study back in the 1940s. Scientists have a far better understanding of the causes of heart disease, and death rates have plummeted. "There has been nothing short of a cardiac revolution," says Michael Lauer, director of the division of cardiovascular sciences at the National Heart, Lung, and Blood Institute. "It's one of the great triumphs of modern medical science." But the precise manner in which our hearts can betray us has not yet been fully revealed. ■

Cassandra Willyard *is a freelance science writer based in Brooklyn, New York.*

1. Sacks, F. M. *et al. Circulation* **102,** 1886–1892 (2000).
2. Jensen, M. K. *et al. J. Am. Heart Assoc.* **1,** jah3-e000232 (2012).
3. Voight, B. F. *et al. Lancet* **380,** 572–580 (2012).
4. Dutta, P. *et al. Nature* **487,** 325–329 (2012).
5. Wang, Z. *et al. Nature* **472,** 57–63 (2011).
6. Petersen, T. H. *et al. Science* **329,** 538–541 (2010).

XBIOTECH

XBiotech researchers manufacture a monoclonal antibody designed to reduce cardiovascular events.

DRUGS

Blood battles

The standard medications for hypertension and cholesterol have lingering issues, but new drugs hold promise for high-risk patients.

BY KATHARINE GAMMON

Everywhere you look in the United States, drugs for high blood pressure and high cholesterol stare back at you — from billboards and television screens to the pages of magazines. Together, these medications account for a global market worth US$75 billion. Moreover, statins, a mere two decades after coming to market, are the best-selling drugs on the planet. Researchers have raised questions about the effectiveness of prescribing drugs to lower cholesterol and blood pressure before a person has heart disease, and nagging problems with side effects still linger. At the same time, new drugs that use novel approaches to treat the same illnesses are in development.

The World Health Organization (WHO) estimates that a third of adults worldwide have raised blood pressure, a condition that causes around half of all deaths from stroke and heart disease. Hypertension costs the US healthcare system $131 million each year. And it is a major cause of cardiovascular disease, the leading cause of death in the United States.

Medications can lower blood pressure in a variety of ways: diuretics squeeze salt and water out of the blood; beta-blockers reduce heart rate; angiotensin-converting enzyme (ACE) inhibitors slow the body's production of angiotensin, thereby relaxing blood vessels and reducing blood pressure; calcium-channel blockers disrupt calcium from entering the smooth muscle cells of the heart and arteries, softening heart contractions and easing the flow of blood.

Some researchers are looking to shore up our fundamental understanding of how blood pressure works in the body. "Blood pressure is a complicated system," says James Wright, a clinical pharmacologist at the University of British Columbia in Vancouver. "When drugs work to lower pressure, we still are figuring out how low pressure can go, and how the drugs impact the peaks and the troughs of blood pressure."

➲ NATURE.COM
Visit *Nature Reviews Cardiology* for the latest research:
nature.com/nrcardio

Wright says that researchers need to know whether antihypertensive drugs work by lowering the maximum blood pressure (systolic) during a heartbeat or the minimum (diastolic), or the difference between the two (pulse pressure). Whether other factors, such as the variability in blood pressure, are involved is not known. Wright is reviewing different classes of antihypertensives to assess the effects on the various forms of blood pressure. Likewise, Francois Gueyffier, a doctor at the Hospices Civils de Lyon in France, suggests a reassessment of blood pressure targets for the elderly, because blood pressure is much more variable than initially thought.

DRUG DIRECTIONS

Pharmaceutical companies are looking for ways to solve the fundamental problems. Monoclonal antibodies may be best known for treating arthritis, but they are now being applied to vascular health. XBiotech, a pharmaceutical company based in Austin, Texas, is using monoclonal antibodies to treat restenosis — a recurrence of the narrowing of a vessel after it's been opened by a stent or surgery — and vascular disease in general.

Monoclonal antibodies are a different way of dealing with that issue, says Hosam El-Sayed, a cardiovascular surgeon at the Methodist Hospital System in Houston, Texas, who has collaborated with XBiotech. In autumn 2012, XBiotech finished a phase II trial of the drug MABp1, which blocks the activity of interleukin-1α, interrupting inflammation related to heart disease. XBiotech announced results from 42 patients showing that the group treated with MABp1 had a 58% reduction in major cardiovascular events.

Most antibodies use mouse cells to create their biologic drugs. Michael Stecher, XBiotech's medical director, says the company is cloning human antibodies instead of using mouse components, which he says makes the drugs safer and more effective.

Instead of creating new drugs, some companies are looking to combine existing medications in innovative ways. Clinical trials have started on a 'polypill' developed by the Indian pharmaceutical company Cipla, based in Mumbai, including three blood-pressure-lowering drugs and a cholesterol-lowering statin. In a 3-month trial of 84 men and women over 50 years of age at Queen Mary, University of London, the pill cut blood pressure by 12% and reduced levels of low-density lipoprotein (LDL) — the so-called 'bad cholesterol' — by 39% (ref. 1).

CHOLESTEROL MEDICATIONS

Several different types of drug can control cholesterol, but statins are the most popular.

WAN HE, MARK N. MUENCHRATH AND PAUL KOWAL, US CENSUS BUREAU.

Statins inhibit the enzyme HMG-coA reductase, thereby blocking the production of cholesterol in the liver. They lower LDLs and raise the levels of 'good cholesterol', high-density lipoproteins (HDLs).

Statins undeniably help reduce deaths from heart disease in high-risk populations. One study followed about 20,000 high-risk patients for 11 years and found that patients who took a statin for 5 years had 23% fewer major vascular events than a control group[2]. In people who have had heart attacks, those who took statins were 18–19% less likely to suffer a combination of heart problems, including strokes, heart attacks or a heart disease-related death[3].

But these drugs are not without risk. Roughly 10% of people who are prescribed statins encounter serious side effects, according to recent research. These range from muscle and joint aches in 5–10% of patients, to moderately elevated blood levels of muscle enzymes in 1–2% of patients, and severe muscle injury and even kidney failure, which happens in 1 in 100,000 patients (0.001%), according to Robert Hegele, an endocrinologist who studies statins at Western University in London, Ontario. "The benefits of statins in terms of preventing heart attacks and strokes, prolonging life and keeping patients with cardiovascular disease out of the hospital dwarf these small risks, but the side effects are a barrier to treatment for some patients," says Hegele.

Hegele says that although statins have made an impact since the 1980s, there is still an unmet need for a way to lower cholesterol. "Some patients fail to reach their targets even with various drug combinations, and patients who cannot tolerate a statin represent yet another group for whom new therapies are needed," he says.

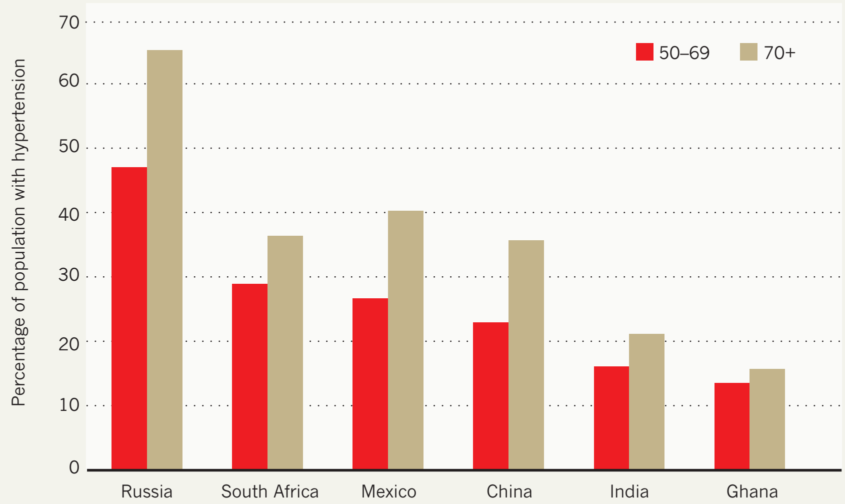
Statins have an undisputed place in treating people with unhealthy levels of cholesterol — generally starting at total cholesterol of 240 milligrams per decilitre (mg dl^{-1}), or an LDL level of 130 mg dl^{-1}. But the medical community has long debated whether statins should be used as a prophylactic. Cardiologist Rita Redberg at the University of California, San Francisco, sees no evidence that statins help healthy people[4]. "Statins are widely used in primary prevention, but the benefits are not clear," says Redberg. "So you're talking about prescribing a drug that doesn't benefit quality of life or living longer." The most optimistic projections suggest that for every 100 healthy people who take statins for 5 years, one or two will avoid a heart attack — but one person will also develop diabetes, says Redberg.

The need to figure out why statins don't work for everyone is motivating some researchers to study what cholesterol does in the body. Cholesterol is an essential part of the human cell; it is involved in everything from the brain to hormones. "Cholesterol is what distinguishes animals from plants. It gives us a nervous system and a brain," says Stephanie Seneff, an artificial-intelligence researcher at the Massachusetts Institute of Technology in Cambridge, who has recently turned her attention to statins.

Seneff says that some of the other side effects of statins, such as muscle weakness and memory loss, are easily put down to the effects of ageing, and might not be recognized as consequences of taking the drug. "Many people are feeling like they're getting old, and they don't realize that the statin drugs are making it worse," says Seneff.

HIGH LEVELS OF HYPERTENSION

For people who are 50-69 years old (red), one-quarter or more suffer hypertension in many countries, and for people who are 70 or more years old (brown) the percentages surpass one-third in many countries, reaching roughly two-thirds in Russia.



REPLACING STATINS

With the widespread prescription of statins, researchers are looking to find alternatives. There are several types of treatment under development to lower LDL cholesterol that work differently to statins.

One of those drugs is a monoclonal antibody that works against an enzyme called PCSK9, which destroys a receptor for LDLs. The drug, called AMG 145, was tested in a recent controlled study in 631 patients aged 18–80 years[5]. In patients who received the shot each month, the drug reduced LDL cholesterol by 42–50% at the end of 12 weeks compared to placebo.

"The idea is: the enemy of your enemy is your friend," says Robert Giugliano, study author and cardiologist at Brigham and Women's Hospital in Boston, Massachusetts. Amgen, the drug's developer, is planning a large phase III study in 20,000 patients.

Another cholesterol-lowering class of drugs in development acts by blocking the action of molecules in the liver that shuttle cholesterol. Four such cholesterylester transfer protein (CETP) inhibitors are in the pipeline; the one furthest along is a pill from Merck called Anacetrapib. Merck is now recruiting 30,000 high-risk patients in Europe, China and the United States for a 4-year trial of anacetrapib called the Randomized Evaluation of the Effects of Anacetrapib through Lipid Modification (REVEAL) trial. The drug will be given in conjunction with a statin to further lower LDL. Anacetrapib has been shown in previous studies to raise HDL by 140% and reduce LDL by 25–40%, says Martin Landray a clinical investigator at Oxford University, UK, who is also involved in REVEAL.

Giugliano stresses that most of the new drugs are for high-risk patients or those who have had a heart attack, and that most people with high blood pressure or high cholesterol should still start their path to wellbeing with a healthy lifestyle. "We're doing better with smoking and hypertension, but weight and exercise is going in the wrong direction. That's the foundation, and the next step will be extra help for people who need help with lowering cholesterol."

Despite the growing list of powerful pharmaceuticals to treat high blood pressure and cholesterol problems, and the ubiquitous adverts proclaiming their success, the most effective treatment might not be found in a pill, but in a pair of running shoes and a salad. ■

Katharine Gammon *is a freelance science writer in Santa Monica, California.*

1. Wald, D. S. *et al. PLoS ONE* **7,** e41297 (2012).
2. Heart Protection Study Collaborative Group. *Lancet* **378,** 2013–2020 (2011).
3. Gutierrez, J. *et al. Arch. Intern. Med.* **172,** 909–919 (2012).
4. Redberg, R .F. & Katz, M. H. *J. Am. Med. Assoc.* **307,** 1491–1492 (2012).
5. Giugliano, R. P. *et al. Lancet* **380,** 2007–2017 (2012).